

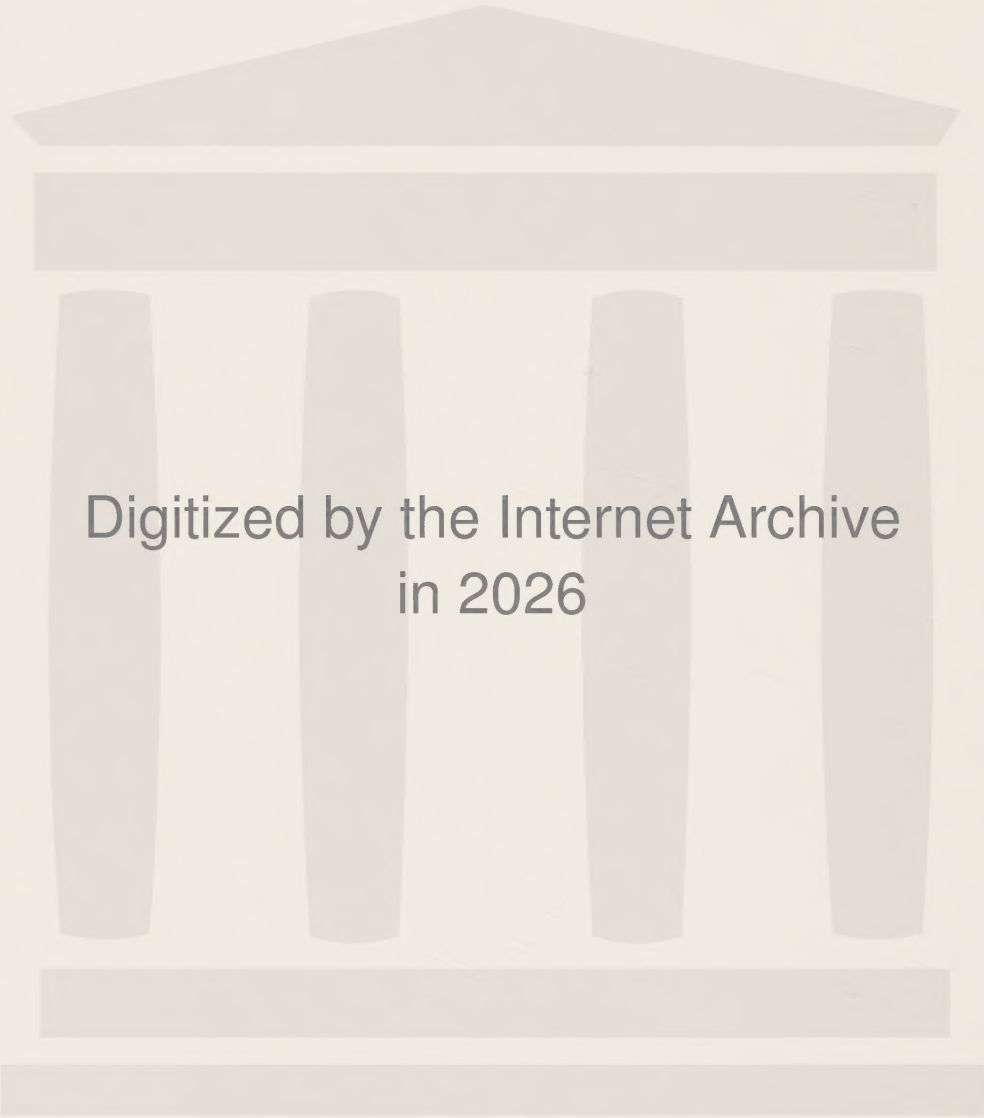
DIE FEHLERBEREINIGUNG BEI VOLKSZÄHLUNGEN UNTER
BESONDERER BERÜCKSICHTIGUNG AUTOMATISCHER KONTROLLEN
UND KORREKTUREN

INAUGURALDISSERTATION

zur Erlangung der Würde eines
Doktors der Wirtschaftswissenschaft
der Abteilung für Wirtschaftswissenschaft
der Ruhr-Universität Bochum

vorgelegt von
Diplom-Volkswirt Harald Rönsch
aus Großrückerswalde/Erzgeb.

1977



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41546602_0063

Die Fehlerbereinigung bei Volkszählungen unter
besonderer Berücksichtigung automatischer Kontrollen
und Korrekturen

INAUGURALDISSERTATION

zur Erlangung der Würde eines
Doktors der Wirtschaftswissenschaft
der Abteilung für Wirtschaftswissenschaft
der Ruhr-Universität Bochum

vorgelegt von
Diplom-Volkswirt Harald Rönsch
aus Großbrückerswalde/Erzgeb.

1977



Referent: Prof. Dr. Klemmer
Korreferent: Prof. Dr. Stier
Dekan: Prof. Dr. Chmielewicz
Tag der mdl. Prüfung: 14. Juli 1977

Inhalt

	Seite
1 Einleitung	5
2 Fehler in Statistiken aus Volkszählungen	9
21 Systematischer Fehler (Bias)	10
22 Zufallsfehler	12
23 Gesamtfehler	13
231 Qualitätsindex	15
232 Qualitätsstandard	17
24 Ausgangssituation für die automatische Fehlerbereinigung	19
3 Fehleridentifikation	21
31 Begriffsbestimmungen	21
32 Wahrscheinlichkeitskontrollen	23
321 Parametrische Verfahren	24
3211 Univariante Verfahren	24
32111 Gross Error-Methode	24
32112 FUNOR (Full Normal Rejection)	26
32113 Annahme-Kontrolle	28
3212 Multivariate Verfahren	29
32121 Bivariater Fall	29
32122 Hauptkomponentenlösung	31
322 Sonstige parametrische Verfahren	35
3221 Anpassung an das Elderton-Pearson'sche System von Häufigkeitsverteilungen	35
3222 Kontrolle auf der Grundlage des Bayes'schen Theorems	36
323 Nichtparametrische Verfahren	38
3231 Order Statistics-Lösung	38
3232 Ungleichungen vom Tschebyscheff-Typ	39
33 Deterministische Kontrollen	41
331 Funktionale Kontrolle	42
3311 Additions- bzw. Nullkontrolle	42
3312 Quotienten-Kontrolle	43
3313 Anwendung funktionaler Kontrollen auf Volkszählungen	44
33131 Art der Merkmale bei Volkszählungen	44
33132 Informationsredundanz	44

332	Automatische Signierkontrollen	47
3321	Kombinationskontrollen	49
3322	Binärkettentechnik (binary string)	49
3323	Zukunft automatischer Signierkontrollen	50
333	Mengenansatz von Fellegi/Holt	51
334	Konstruktionsprinzipien des Kontrollrahmens	53
34	Statistische Qualitätskontrolle	55
341	Prozeßkontrolle	56
342	Annahmestichproben	57
343	Hybride Verfahren	61
35	Automatische Bestandskontrollen	63
351	Bevölkerungsbegriff in deutschen Volkszählungen	63
352	Ordnungssystem	64
353	Paarigkeitskontrolle von Ordnungsbegriffen (Dateienvergleich)	67
354	Kleinräumige Gliederung	68
355	Verbesserung der Erfassungsgenauigkeit	69
356	Erfahrungen	71
4	Fehlerallokation	73
41	Einzelfehler	73
42	Unabhängigkeitsmodell	75
43	Zuverlässigkeitsmaß	76
44	Ableitung von "Erklärungsmengen" aus deterministischen Kontrollen	79
5	Fehlerbeseitigung	82
51	Bestimmung wichtiger Begriffe	82
52	Parameterschätzung	83
521	Schätzung von Antwortwahrscheinlichkeiten	83
522	Maximum Likelihood-Schätzungen	85
523	Eignung für die Volkszählung	90
53	Einzelfall-Korrekturen	92
531	Stochastische Korrekturen	92
5311	Doppeln	92
5312	Cold Deck	95
5313	Hot Deck	97
5314	Monte Carlo-Methode	100

532	Deterministische Korrekturen	103
5321	FUNOM (Full Normal Modification)	103
5322	Funktionale Korrekturen	103
5323	Korrektur von Signierbegriffen	108
53231	Einzelfehler-Korrekturcode	108
53232	Beispiele für deterministische Lösungen aus Volkszählungen	109
6	Analyse des Systems der Fehlerbereinigung bei der Volks- zählung 1970 in der Bundesrepublik	116
61	Ablaufplan	116
62	Bestimmung der Wohnbevölkerung	122
63	Qualitätskontrolle der automatischen Korrekturen	129
7	Bewertung der automatischen Fehlerbereinigung	133
71	Vorteile gegenüber manuellen Verfahren	133
72	Grenzen der Anwendbarkeit	135
8	Bedeutung der Fehlerforschung	140
81	Erfassungsfehler	140
82	Totaler Informationsausfall	142
83	Antwortfehler	143
84	Lesefehler	145
9	Schlußbetrachtung	151
	Literaturverzeichnis	155
	Anhang	165

1 Einleitung

Anlaß für diese Untersuchung war die Mitarbeit des Verfassers an der Volkszählung 1970 in der Bundesrepublik. Da bei statistischen Vorhaben - besonders dieser Größenordnung - auf vielen Fragebögen Informationen fehlen oder falsch sind, andererseits die statistische Gesamtheit in den Ergebnistabellen in allen Teilgesamtheiten vollständig nachgewiesen wird, stellt sich dem Verfasser die Frage nach der Art und Weise der Fehlerbereinigung. Aufschlüsse erwartet er durch die Analyse der Arbeitsgänge, in denen die Fehler identifiziert und beseitigt werden. Dabei ist zu prüfen, welche Auswirkungen sich auf die Verteilungen der bei Volkszählungen erhobenen Merkmale ergeben.

Daß der Computer für die zeit- und kostensparende Verarbeitung großer Datenmengen geeignet ist, kann vorausgesetzt werden. Schließlich war er geradezu für den Einsatz bei Volkszählungen entwickelt worden.¹⁾ Daß er Kontrollen schnell und konsistent ausführen kann, war bereits bei der Zählung 1961 erprobt worden. 1970 wurde nun erstmals versucht, den Rationalisierungsvorteil der automatischen Datenverarbeitung ganz zu nutzen und auch die Fehlerbeseitigung in die Maschinenprogramme zu integrieren. Daß durch den Wegfall umfangreicher Korrekturarbeiten Arbeitskräfte eingespart werden können, ist offensichtlich. Außerdem begründen die Konsumenten mit dem Einsatz leistungsfähiger Computer ihren Anspruch auf erste Informationen oft schon kurze Zeit nach dem Stichtag.²⁾ Dabei wollen sie auf die Zuverlässigkeit der Daten nicht verzichten. Der Verfasser räumt zwar ein, daß angesichts der Terminzwänge, die bei einem so bedeutsamen Projekt herrschen, Zugeständnisse an die Qualität gemacht werden müssen; andererseits hat ein Teil der Ergebnisse bei der gegenwärtigen Frequenz von Volkszählungen in vielen Bereichen den Charakter von Eckwerten für zehn Jahre. Da von Totalzählungen außerdem der Nachweis regional und sachlich tief gegliederter Ergebnisse erwartet wird, müssen die Anforderungen an die Genauigkeit hoch sein. Der Ver-

1) Hollerith entwickelte für die Aufbereitung des amerikanischen Census 1890 eine elektrische Zählmaschine, die Merkmale in Form von gelochten Karten zählen konnte.

2) Vgl. Fürst, G., Maschinenverwendung und Automatisierung in der Statistik, in: Allgemeines Statistisches Archiv, Bd 43, H. 4, München 1959, S.313.

fasser stimmt der Ansicht Morgensterns zu, daß sich die Fehlerbehandlung in den Sozialwissenschaften an den Naturwissenschaften orientieren sollte.¹⁾

Die Wirksamkeit des Computers als zentrales Instrument der Datenbereinigung kann aber nur dann richtig beurteilt werden, wenn die Arten der Fehler, ihr Zustandekommen und die Bemühungen zu ihrer Beseitigung bereits im Vorfeld der Datenverarbeitung in die Betrachtung einbezogen werden. Die technisch-organisatorische Absicht muß es sein, die Maschinenprogramme so zu gestalten, daß manuelle Eingriffe in den Ablauf unterbleiben. Es ist zu prüfen, ob die methodischen Grundlagen ausreichen, um dieses Ziel ohne Qualitätsverluste zu erreichen. Da Fehlerbereinigung der Erfolgskontrolle bedarf, sind Genauigkeitsmessungen durchzuführen oder - soweit sie vorliegen - kritisch zu würdigen.

Der Verfasser hat die zur Durchführung der Volkszählung 1970 ergangenen Arbeitsanweisungen des Statistischen Bundesamtes, die bei vorhandenem Interesse auch der Öffentlichkeit zur Verfügung stehen, auf ihre Auswirkungen zur Qualitätsfindung untersucht. Auf eine offizielle Dokumentation der methodischen Voraussetzungen der Zählung 1970 konnte er nicht zurückgreifen. Das erschwerte die Arbeit sehr, hatte aber den Vorteil, daß Verbesserungsvorschläge auch für einzelne Vorgänge im Projektablauf entwickelt werden konnten, denen sonst möglicherweise unkritisch zugestimmt worden wäre. In USA lagen Erfahrungsberichte zum Census 1970 ab 1973 vor. Außerdem existiert in USA, Kanada und den skandinavischen Ländern eine umfangreiche Literatur zu Fragen der Fehlerbereinigung bei statistischen Erhebungen. Auf diese Literatur, die mit vorwiegend praktischen Bezügen erst seit der 35. Sitzung des Internationalen Statistischen Instituts in Belgrad 1966 (mit den Arbeiten von Pritzker, Ogus, Hansen sowie Nordbotten und Szameitat, Zindler) verstärkt auftrat, konnte bei der Analyse der Fehler, ihrer Kontrollen und Korrekturen eines so vielschichtigen und komplizierten Unternehmens wie einer Volkszählung zurückgegriffen werden.

1) Morgenstern, O., Über die Genauigkeit wirtschaftlicher Beobachtungen, 2. erw. Aufl., Physica, Wien-Würzburg 1965, S.4 und 48.

In der deutschen Literatur wurde die Fehlerbereinigung in der Statistik - soweit sie operativ ist, also das Zustandekommen der statistischen Ergebnisse betrifft - bisher nicht umfassend behandelt. Weichselberger hat 1960 über Ergebniskontrollen bei Volkszählungen - und da besonders im Hinblick auf die Erfassungsgenauigkeit der Wohnbevölkerung und ihrer Altersverteilung - gearbeitet.¹⁾ Damals wurde der maschinellen Datenbehandlung noch kein größeres Gewicht beigemessen. Der Verfasser hält eine gründliche Analyse der Fehlerbereinigung im Aufbereitungsprozeß jedoch für überaus wichtig, weil nur hier die Verlässlichkeit der statistischen Ergebnisse unmittelbar beeinflußt werden kann. Nachträglich kann die Genauigkeit lediglich noch konstatiert werden.

Modellbetrachtungen könnten bei dieser umfassenden Themenstellung nur Teilaspekte (einzelne Fehlerarten, einzelne Kontroll- und Korrekturtypen) einbeziehen. Simulationsstudien z.B., auch wenn sie sich auf wenige Alternativen der Fehlerbereinigung beschränken, erweisen sich als sehr aufwendig. Einzelne dürften kaum in der Lage sein, originäre empirische Studien zu finanzieren, vor allem nicht, wenn sie keinen Zugriff zu einer Datenverarbeitungsanlage und einen Stab von Mitarbeitern haben. Die Studien des Verfassers über die Fehler der optischen Beleglesung²⁾ zeigten die Grenzen des Erreichbaren bereits deutlich auf. Daher muß die Methode der vorliegenden Untersuchung darin bestehen, auf der Basis der Analyse der theoretisch entwickelten, nur zum Teil praktisch erprobten Verfahren zur Fehleridentifikation, Fehlerallokation und Fehlerbeseitigung, Vorschläge für die Anwendung auf Volkszählungen zu erarbeiten. Für erfolgversprechende Verfahren, über die keine praktischen Erkenntnisse vorliegen, muß die Durchführung größerer Forschungsvorhaben vorgeschlagen werden. Das umfassende Konzept der Arbeit - wenngleich es die deskriptiven Verfahren nachträglicher Fehlerforschung weitgehend ausklammert - erscheint dem Verfasser richtig und notwendig, solange in der deutschen Literatur ein grundlegendes Werk fehlt. Hier hat sich der Verfasser bemüht, ein gültiges begriffliches Instrumentarium zu schaffen.

1) Weichselberger, K., Kontrollen der Ergebnisse von Volkszählungen, Habil.-Schrift, Köln 1962.

2) Vgl. S.147 ff.

Die Vorschläge zur Qualitätsverbesserung dienen vermutlich primär dem Produzenten von Statistiken aus Volkszählungen. Der Verfasser weiß jedoch aus seiner Beratungstätigkeit für die Konsumenten, daß diese großen Wert auf Informationen über die Art der Fehlerbereinigung bei Volkszählungen legen. Aus ihrer Unkenntnis des Arbeitsablaufs zwischen der Ausfüllung der Belege und der Veröffentlichung der Ergebnisse resultiert auch ihre Überraschung, wenn sie erfahren, daß zum Zeitpunkt ihrer ersten Anfrage nach Ergebnissen größere Städte noch Nacherhebungen durchführen, also mit der zentralen Bearbeitung der Belege für diese Gemeinden nicht einmal begonnen werden konnte. Entscheidender ist natürlich, daß die Konsumenten Kriterien für die Bewertung der Ergebnisse erhalten. Dieses Anliegen wird alljährlich auf den Tagungen der Deutschen Statistischen Gesellschaft (Statistische Woche) vorgetragen.

2 Fehler in Statistiken aus Volkszählungen

Die Volkszählung als das Kernstück der Bevölkerungsstatistik gibt Aufschluß über die Bevölkerungszahl und die regionale Verteilung der Bevölkerung an einem bestimmten Tag (Zählungstichtag) und darüber hinaus über ihre demographische und soziale Struktur sowie ihre wirtschaftliche Betätigung.

Volkszählungen gehören zu den umfangreichsten Vorhaben der amtlichen Statistik, wenn sie als Totalstatistiken durchgeführt werden, also die gesamte Bevölkerung umfassen.

Wegen ihrer 10jährigen Periodizität¹⁾ weisen Volkszählungen im Vergleich zu laufenden Statistiken organisatorische Besonderheiten auf, die beträchtlichen Einfluß auf die Genauigkeit der Zählungsergebnisse haben können. So steht in den in der Bundesrepublik für die Datensammlung verantwortlichen Gemeinden ein Zählungsleiter selten für zwei Zählungen zur Verfügung; während sich die Statistischen Landesämter bei ständig wiederkehrenden Statistiken auf einen eingearbeiteten Mitarbeiterstab stützen können, sind sie gezwungen, für die Aufbereitung einer Volkszählung in großem Umfang neue Arbeitskräfte einzustellen, die weder Fachwissen noch Erfahrungen für die spezifischen Arbeitsgänge mitbringen.

Die Beteiligung vieler Personen an dem statistischen Projekt "Volkszählung" macht dieses einerseits besonders anfällig für das Auftreten von Fehlern, andererseits gestattet die Volkszählung als Totalstatistik den Nachweis von Ergebnissen in tiefer regionaler und sachlicher Gliederung, wodurch alle Mängel im Arbeitsablauf schonungslos aufgedeckt werden. Auf den

1) Die Verfassung der USA vom 17. Sept. 1787 sieht vor, daß zur Aufteilung der Wahlmandate auf die Bundesstaaten eine Zählung der Bevölkerung stattfinden muß, die sich im Abstand von zehn Jahren wiederholt. Da ein Weltzensus internationale Vergleiche wirtschaftlicher und sozialer Entwicklungen ermöglicht, sollen auf Empfehlung der UN Volkszählungen weltweit in Jahren mit der Endziffer "0" durchgeführt werden; Principles and Recommendations for National Population Censuses, in: Statistical Papers, Ser. M, No. 27, Hrsg. United Nations, New York 1958, S.3.

nächsten Seiten werden die Fehlerarten, der Nutzen ihrer Messung und die Voraussetzungen für ihre automatische Behandlung untersucht.

21 Systematischer Fehler (Bias)

Dieser Fehler tritt bei jeder praktischen Durchführung einer Statistik auf, gleichgültig, ob es sich um eine Total- oder Repräsentativstatistik handelt. Deming hält eine vollkommene Zählung für einen Mythos.¹⁾

Das Wesen systematischer Fehler besteht darin, daß sie keine Ausgleichstendenzen bei Vergrößerung des Datenmaterials aufweisen, sondern die Ergebnisse konsistent in einer Richtung verfälschen. In der Stichprobentheorie wird der systematische Fehler als Nicht-Erwartungstreue gekennzeichnet, d.h. es besteht eine Differenz zwischen dem Erwartungswert einer Schätzfunktion und dem unbekannten ("wahren") Wert der Grundgesamtheit. Der im Angelsächsischen auch verwendete Begriff "non-sampling error"²⁾ drückt aus, daß der Fehler keinen Zufallsschwankungen unterworfen ist. Bei großer Sorgfalt kann er zwar gering gehalten, aber dennoch nicht vermieden werden.

In einem kurzgefaßten Überblick werden wichtige Quellen systematischer Fehler in den einzelnen Phasen des Entstehungsprozesses von Statistiken aus Volkszählungen genannt.

Fehler können bereits bei unsachgemäßer Definition der statistischen Gesamtheit, der Merkmale und Merkmalsausprägungen entstehen. Sie können auftreten, wenn die zeitliche, regionale und sachliche Zuordnung von statistischen Einheiten zu bestimmten

1) Deming, W. E., Some Theory of Sampling, Wiley & Sons, Chapman & Hall, New York/London 1950, S.24.

2) Vgl. Hansen, M. H., Waksberg, J., Research on Non-Sampling Errors in Censuses and Surveys, in: Review of the International Statistical Institute, Vol. 38, No. 3, The Hague 1970, S.317-332.

theoretischen Konzepten mißlingt.¹⁾ Ein abstraktes Untersuchungsziel kann möglicherweise nur durch Vergrößerung im Zählvorgang realisiert werden (Adäquationsproblem).

Die Erhebungstechnik kann Ursache für weitere Fehler sein. Der Einsatz von Interviewern kann Verfälschungen mit sich bringen dadurch, daß bestimmte Fragen uneinheitlich interpretiert oder Suggestivfragen gestellt werden. Aber auch die Einstellung der Auskunftspersonen zur statistischen Erhebung, ihr Erinnerungsvermögen und ihre Haltung zu "Intimfragen" (z.B. Einkommen, Gesundheit, uneheliche Kinder)²⁾ können zu Mängeln führen. Art, Gestaltung und Begriffsbildung des Fragebogens sind weitere schwache Stellen. Die Verwendung von Begriffen wie "überwiegend" oder "hauptsächlich" kann die Auskunftsperson vor Entscheidungsprobleme stellen, denen sie durch Mehrfach-Antworten ausweicht. Ein maschinell lesbarer Beleg, auf dem weitgehend vorgegebene Antworten einfach zu markieren sind, kann die Befragten zu oberflächlicher Handhabung verleiten.

Es ist gesetzlich vorgeschrieben, daß die Auskunftspflichtigen die bei einer Volkszählung an sie gerichteten Fragen wahrheitsgemäß und vollständig beantworten. Dieses staatliche Interesse wird durch Geldbußen geschützt.³⁾ In der Praxis können nur Auskunftsverweigerungen im Sinne einer Rückweisung des gesamten Fragenkatalogs geahndet werden. Außerdem kommt es zu Untererfassungen als Folge von Mängeln in der Erhebungsorganisation. Nicht gezählte Personen bewirken einen totalen Informationsausfall. Die unvollständige Beantwortung eines Fragebogens erzeugt einen teilweisen Informationsausfall.

1) Vgl. Pfanzagl, J., Artikel "Fehlertheorie" in: Handwörterbuch der Sozialwissenschaften (HdSW), Bd 3, Fischer, Stuttgart, Mohr, Tübingen, Vandenhoeck & Ruprecht, Göttingen 1961, S.482.

2) Zu dem Problem, ob einzelne Fragen einen Eingriff in die Intimsphäre darstellen vgl. Mündliche Anfrage des Abg. Strohmayer betr. Repräsentativbefragung in der Volkszählung 1970, Bundestags-Drucksache VI/415 - Schriftliche Antwort 35. Sitzung, S.1716 A; BVerfGE 27, S.1 ff. (S.6/7).

3) In der Bundesrepublik § 10 Abs. 2 und § 14 Gesetz über die Statistik für Bundeszwecke (StatGes) vom 3. September 1953 (BGBl. I S.1314).

Totaler und teilweiser Informationsausfall spielen als Non Response-Problem in der Fehlerbereinigung eine wichtige Rolle.

Aufbereitungsfehler entstehen im Falle von Volkszählungen bei den Trägern der amtlichen Statistik (Statistische Landesämter in der Bunderepublik, Bureau of the Census in USA) nach Eingang der Belege. Sie betreffen die manuellen Vollständigkeits- und Vollzähligkeitskontrollen, das Signieren verbaler Informationen auf den Fragebögen, die Datenerfassung mittels Lochung oder Beleglesung, aber auch den automatischen Kontroll- und Korrekturprozeß. Gerade die Phase der zentralen automatischen Fehlerbereinigung, in der Informationsausfälle und Informationsverfälschungen erkannt und beseitigt werden sollen, kann sich als Quelle für neue systematische Fehler erweisen.

Darstellungs- und Interpretationsfehler sind der Vollständigkeit halber noch zu erwähnen. Tabellarische und graphische Darstellungen weisen manchmal Mängel auf. Fehler bei der Interpretation bzw. Analyse statistischer Ergebnisse sind nicht selten auf die vom Erhebungsziel der Statistik abweichende Verwendung durch den Konsumenten zurückzuführen. Da beide Fehlerarten mit dem Zustandekommen der statistischen Ergebnisse nichts zu tun haben, wird hier nicht näher auf sie eingegangen.

22 Zufallsfehler

Bei einer Volkszählung, die als Totalstatistik durchgeführt wird, kann definitionsgemäß kein Zufallsfehler vorkommen, wenn dieser als Stichprobenfehler verstanden wird. Stichprobenfehler treten auf, wenn nicht alle Einheiten der zu untersuchenden statistischen Gesamtheit für die Statistik herangezogen werden, sondern nur nach dem Zufallsprinzip ausgewählte Einheiten. Die praktische Durchführung der Auswahl kann dabei zu einem systematischen Auswahlfehler führen. Das Wesen des Zufallsfehlers besteht darin, daß er Ausgleichstendenzen bei Vergrößerung des Datenmaterials aufweist.

Es besteht allerdings auch die Auffassung, die Ergebnisse sich wiederholender Totalzählungen als Abfolge von Realisationen von Zufallsvariablen und somit als stochastischen Prozeß anzusehen. Dieser Aspekt wird im allgemeinen vernachlässigt, wenn die Ergebnisse einer einzelnen Zählung zur Beschreibung von Massenerscheinungen verwendet werden.¹⁾

Unter den im vorigen Abschnitt aufgeführten systematischen Fehlern kann es Fehlerkomponenten geben, die zufallsähnlichen Charakter besitzen. So kann z.B. der Antwortfehler neben der systematischen eine Zufallskomponente, die Antwortvarianz, enthalten. Ein Teil der Auskunftspersonen kann bei der Frage etwa nach dem Einkommen übertreiben, ein Teil untertreiben. Gewisse Fragen des Fragebogens können Antworten in eine Richtung wie in die Gegenrichtung auslösen. Für größere Gebiete kann der Saldo fast null sein, für kleinere Gebiete, seltenere oder fehleranfälligere Merkmale kann er größer sein. Die Unterscheidung ist bei der Fehlerforschung von Bedeutung.

In zunehmendem Maße werden in der Verknüpfung von Total- und Repräsentativstatistiken Vorteile gesehen. Diese Kombination ist in USA schon seit 1950 üblich, in der Bundesrepublik wurde sie erstmals 1970 hergestellt. Die Gründe liegen darin, daß die Grenzen der Belastbarkeit der Auskunftspersonen als erreicht angesehen werden, so daß ein Teil des Fragenkataloges nur noch an eine Auswahl aus der Bevölkerung gerichtet wird, zum anderen, daß Forschungsergebnisse die Fehlerhaftigkeit von Totalstatistiken bewußt gemacht haben.

23 Gesamtfehler

Der Gesamtfehler kennzeichnet die Genauigkeit oder die Qualität einer Variablen bzw. einer Statistik. In einer kürzlich veröffentlichten Studie wird eine Übersicht über Terminologien verschiedener Statistiker zu Begriffen wie "Genauigkeit",

1) Vgl. Stichproben in der amtlichen Statistik, Hrsg. Statistisches Bundesamt, Wiesbaden, Kohlhammer, Stuttgart-Mainz 1960, S.16 und 49.

"Präzision" und "Zuverlässigkeit" gegeben.¹⁾ Die Übereinstimmungen bei der Definition der "Genauigkeit" sind noch am größten. Allen Erläuterungen gemeinsam ist die "Nähe zum wahren Wert". Cochran spricht von der "Größe der Abweichung vom wahren Mittelwert". Kendall drückt es als "Nähe einer Beobachtung zu der Quantität, die beobachtet werden sollte", aus. Hier ist der mögliche Definitionsfehler mit einbezogen. In diesem Zusammenhang ist darauf hinzuweisen, daß sich der Gesamtfehler in dem Zeitraum zwischen Abschluß der statistischen Erhebung und Verfügbarkeit der Daten, also durch Aktualitätsverlust, verändern kann.²⁾

Der "wahre" Wert kann nur modellhafte Bedeutung haben, da es den Idealfall einer Zählung (Stichprobe oder Totalzählung) ohne jeden Mangel nicht gibt. Immerhin gilt er als existent in dem Sinne, daß er durch unbegrenzten Einsatz an Kosten und Zeit erreichbar wäre. In der Praxis nehmen die Statistischen Ämter durch Wiederholung von Statistiken und Durchführung ähnlicher Statistiken verschiedene Schätzungen der Werte einzelner Variablen vor und ermitteln so mehr intuitiv "wahre" Werte. In derselben Weise wird in der o.g. Studie ein Kompromiß zwischen dem Ideal und dem vernünftigerweise Erreichbaren vorgeschlagen, indem zur Verwendung eines "optimalen" Wertes geraten wird, der aus einem "bevorzugten Verfahren" stammen soll.³⁾ Damit kann offenbar nur eine Stichprobenstatistik gemeint sein, die infolge besonderer Sorgfalt bei der Planung, Erhebung und Aufbereitung Normcharakter trägt.

Wenn man sich zur Berechnung von Gesamtfehlern (z.B. Antwort-

-
- 1) Sanchez-Crespo, J. L., Notes on the Accuracy, Precision and Reliability of Statistical Data, Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975, S. 749-758.
 - 2) Vgl. Szameitat, K., Deininger, R., Some Remarks on the Problem of Errors in Statistical Results, in: Bulletin of the International Statistical Institute, Proceedings of the 36th Session, Vol. XLII, Book 1, Sydney 1967, S. 67/68.
 - 3) Sanchez-Crespo, J. L., Notes on the Accuracy, Precision and Reliability of Statistical Data, a.a.O., S.754.

fehlern) in Totalstatistiken stichprobenhaft durchgeführter Kontrollerhebungen bedient, dann läßt sich der Gesamtfehler wie folgt definieren:

$$(\text{Mittlerer quadratischer Fehler})^2 = (\text{Standardfehler})^2 + (\text{systematischer Fehler})^2 \quad 1)$$

Der Gesamtfehler (z.B. Antwortfehler) besteht also aus der Summe der Varianz (z.B. Antwortvarianz) und dem Quadrat des Bias (z.B. Antwortbias) des Schätzwertes von Totalen bzw. Parametern einer Variablen. Wenn der Gesamtfehler in Stichprobenstatistiken ermittelt werden soll, ist zusätzlich die Stichprobenvarianz zu berücksichtigen.

Es bedarf intensiver Fehlerforschung, um vor allem die systematische Komponente nach Richtung und Größe zu bestimmen. Als Schätzwert für den Antwortbias einer Totalstatistik wird oft die Differenz zwischen dem dort nachgewiesenen Wert und dem "optimalen" Wert einer Kontrollerhebung benutzt. Der Antwortbias repräsentiert allerdings selbst Fehler unterschiedlicher Herkunft.

231 Qualitätsindex

Bei der praktischen Durchführung von Genauigkeitsmessungen bei Volkszählungen ist zu berücksichtigen, daß die statistische Gesamtheit entsprechend den diskreten Ausprägungen der Variablen in Teilmassen zerfällt (z.B. Altersgruppen, sozioökonomische Gruppen), wobei jede Gruppe Klassifikationsfehler aufweisen kann. Der Gesamtfehler für eine einzelne Variable setzt sich also aus Fehlern dieser Teilmassen zusammen. Außerdem besteht ein Veröffentlichungsprogramm oft aus einer Reihe von mehrdimensionalen Tabellen, in denen Merkmalskombinationen mit spezifischer Fehlerstruktur vorkommen. Da sich die Mehrzahl der Konsumenten für einzelne Merkmalsbereiche interessiert, erscheint es wenig sinnvoll, ihnen eine Globalzahl über alle Merkmale hinweg anzubieten, die - ähnlich einem Preisindex - individuelle Phänomene verdeckt.

1) Vgl. Stichproben in der amtlichen Statistik, Hrsg. Statistisches Bundesamt, a.a.O., S.112.

Der Statistik-Produzent bevorzugt zur Verbesserung künftiger Verfahren Fehleranalysen einzelner Arbeitsphasen (z.B. Feldarbeit, Signierprozeß), so daß auch von ihm keine Nachfrage nach einem Qualitätsindex ausgeht. Ein solcher Index wäre auch nur zweckmäßig, wenn er als Vergleichsmaßstab dienen könnte. Das würde eine Konstanz im Aufbau des Fragenkatalogs und des technisch-organisatorischen Ablaufs der Zählung voraussetzen.

Die UN schlägt den Statistischen Ämtern, ausgehend von einer Studie über die Genauigkeit bei der Ermittlung des Brutto-sozialprodukts, vor, anstelle einer einzigen numerischen Größe, einzelne Fehlermaße mit der qualitativen Beschreibung von Schwächen und Stärken der Konzepte (Definition, Erhebungs-, Aufbereitungsmethode) zu kombinieren.¹⁾ Da die Volkszählung als Mehrzweckstatistik die Untersuchung sehr verschiedenartiger Merkmalsbereiche der Bevölkerung zum Ziel hat, erscheint dem Verfasser dieses Vorgehen auch hier von Vorteil.

Qualitätsvergleiche bei einer Zählung über verschiedene Erhebungsjahre auf nationaler Ebene sind schon problematisch genug; wenn sie sich auf mehrere Staaten beziehen, sind die Schwierigkeiten - nicht zuletzt aus Gründen des Prestiges - geradezu unüberwindlich. Zur Verbesserung vor allem der internationalen Vergleichbarkeit der Durchführung von Statistiken wurde die Entwicklung eines Schlüssels vorgeschlagen, der Einblick in die Arbeitsabläufe gewährt.²⁾ Dies kann als Versuch gewertet werden, die qualitative Beschreibung der Konzepte, von der oben die Rede war, zu standardisieren. Natürlich kann die bloße Aufzählung der angewandten Verfahren den Nachweis objektiver Genauigkeitsmessung nicht ersetzen. Bei jedem Arbeitsschritt bleibt die Frage nach der Qualität, mit der dieser ausgeführt wurde.

1) Accuracy and Statistics, Arbeitspapier, bearb. von Statistics Canada, Hrsg. United Nations, Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, 22nd Plenary Session, Genf 1974.

2) Szameitat, K., Deininger, R., Some Remarks on the Problem of Errors in Statistical Results, a.a.O., S. 80 ff; Schlüssel für die Volkszählung 1961 in der Bundesrepublik im Anhang.

232 Qualitätsstandard

Der Qualitätsstandard einer Statistik umfaßt nach Meinung des Verfassers nicht nur bestimmte Mindestanforderungen an die Genauigkeit, sondern auch Ansprüche an den Informationsgehalt und die Schnelligkeit, mit der Ergebnisse vorgelegt werden. Diese Ziele stehen in Konflikt miteinander. Die Aufteilung der Ressourcen ist abhängig von der Wertigkeit, die diesen Zielen beigemessen wird.

Vorausgesetzt, aus übergeordneter politischer Anschauung wird unter Berücksichtigung der Situation der öffentlichen Haushalte die Durchführung einer Statistik überhaupt als notwendig erachtet, kann aus Gründen der Sparsamkeit über die Höhe des Anteils, in dem statistische Einheiten in die Erhebung einzu beziehen sind, sowie über den Umfang des Fragenkatalogs durchaus Uneinigkeit unter den Beteiligten herrschen. Entscheidungen zu diesen Fragen haben direkte Auswirkungen auf den Qualitätsstandard. Es erscheint dringend erforderlich, für die Abstimmung der Reihenfolge der Ziele, die Kosten für alternative Konzepte der Fehlerbereinigung zu spezifizieren.

Im Prinzip sollte bei einer Statistik sogar das Erfordernis einer Fehlerbereinigung geprüft werden. Es mag Fälle geben, in denen der Zählungsplan so gut erfüllt wird, daß Kontrollen und Korrekturen überflüssig erscheinen, oder daß der Informationsbedarf so dringlich ist, daß die Phase der Fehlerbereinigung entfallen muß und ungefähre Aussagen in Kauf genommen werden. Wenn dann auch noch auf die Übereinstimmung der Ergebnisse in Tabellen unterschiedlicher Klassifikation kein Wert gelegt wird, erweist sich eine Fehlerbereinigung tatsächlich als entbehrlich.

Um objektive Kriterien für Auseinandersetzungen um den Qualitätsstandard zu gewinnen, schlägt Nordbotten vor,¹⁾ "Entscheidungsstichproben" (decision sample) aufzubereiten und zu

1) Nordbotten, S., Automatic Editing of Individual Statistical Observations, Statistical Standards and Studies, No. 2, Hrsg. United Nations, Conference of European Statisticians, New York 1963, S.33-44.

analysieren. Diese Stichproben sollen Wissen über die Gesamtheit vermitteln, die Grundlage für die Entscheidung über die Durchführung der Fehlerbereinigung liefern und Kriterien für die Wahl derjenigen Kontroll- und Korrekturverfahren beibringen, mit denen die größte meßbare Qualitätsverbesserung im Sinne einer Reduktion des Gesamtfehlers (oder auch nur der Zufallsfehler) zu erzielen ist. Die Methode, zu der er rät, um den Stichprobenumfang oder - wenn dieser gegeben ist - die Risiken für den Produzenten oder Konsumenten zu bestimmen, ist die in der Statistischen Qualitätskontrolle übliche, Operationscharakteristiken zu berechnen. Neben den beiden Risiken (in der Testtheorie: Fehler 1. und 2. Art), enthalten diese die Gut- und Schlechtgrenzen sowie Stichprobenumfang und Annahmehzahl.¹⁾

Die Qualitätsansprüche, die Produzenten und Konsumenten an Statistiken stellen, sind schon im Falle einer Variablen schwer zu quantifizieren, bei einer Mehrzweckstatistik wie der Volkszählung, bei der wegen der Vielzahl von Variablen Art und Umfang späterer Analysen nicht vorhersehbar sind, hält der Verfasser die Formulierung eines Qualitätsstandards praktisch nicht für möglich. Wegen des traditionell aus Volkszählungen geforderten Nachweises von Ergebnissen in tiefer regionaler und sachlicher Gliederung, ist ein Höchstmaß an Genauigkeit in allen Arbeitsphasen geboten. Da die Tabellen weder Felder "ohne Angabe" enthalten noch Inkonsistenzen in Untergruppen oder Totalen aufweisen sollen, besteht an der Notwendigkeit von Maßnahmen der Fehlerbereinigung kein Zweifel.

In der Bundesrepublik werden vor Volkszählungen Probeerhebungen durchgeführt. Aus den genannten Gründen werden sie nicht dazu benötigt, um über die Fehlerbereinigung als solche zu entscheiden. Sie dienen vielmehr dem Ziel, Fehlerquellen rechtzeitig aufzuspüren, um die Erhebungs- und Aufbereitungskonzepte zu verbessern.

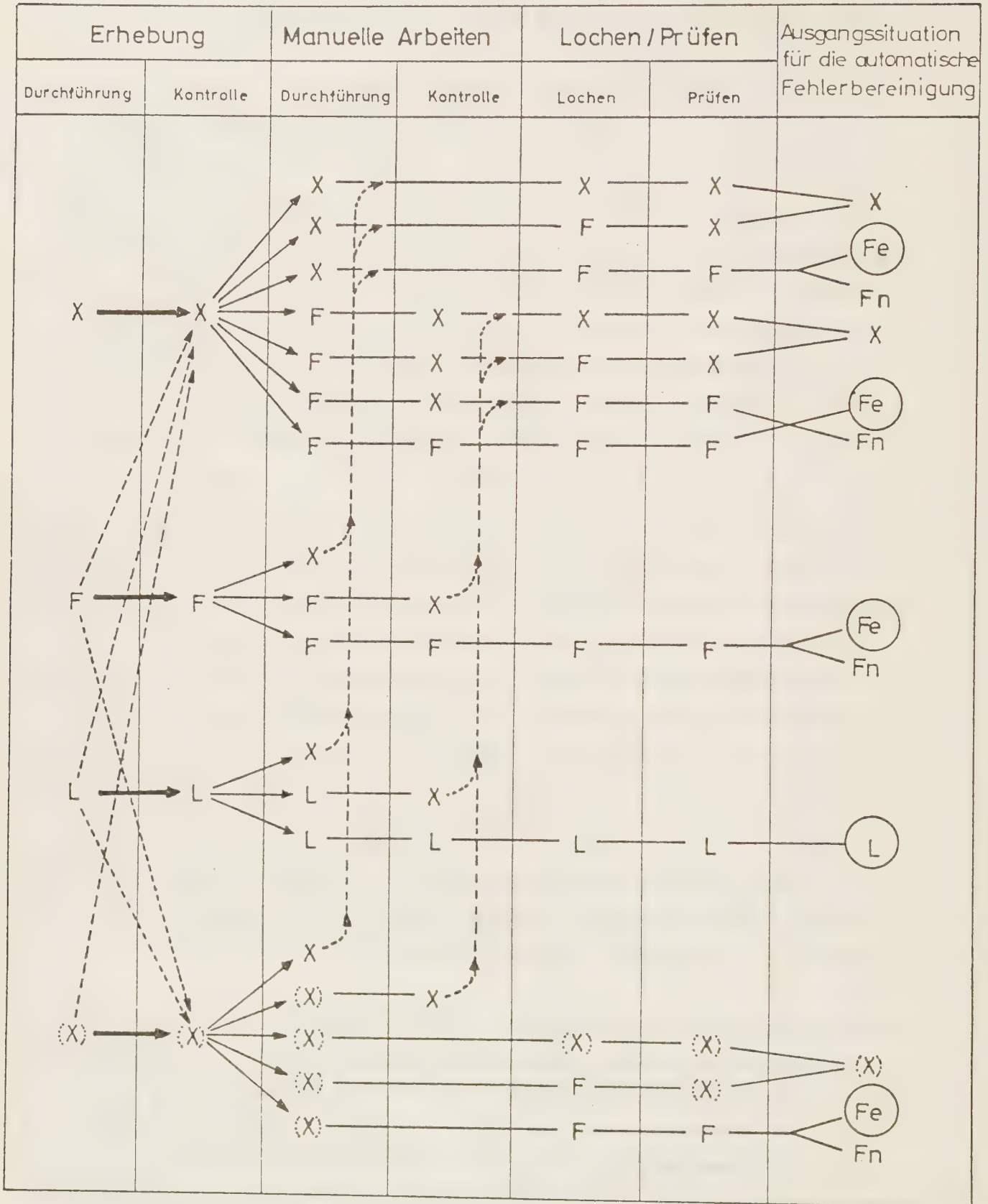
1) Vgl. Abschnitt 34 "Statistische Qualitätskontrolle" S.55 ff.

24 Ausgangssituation für die automatische Fehlerbereinigung

Die Grafik auf der nächsten Seite gibt einen schematischen Überblick darüber, wie Fehler in den wichtigsten Arbeitsphasen der Volkszählung auftreten, wie sie identifiziert und beseitigt werden können oder die Ergebnisse schließlich beeinflussen. Ein Beispiel: Eine Angabe sei richtig erhoben und in der Erhebungskontrolle unbeanstandet geblieben. Wenn es sich bei der Angabe um eine verbale Information handelt, die mit Hilfe eines Schlüsselsverzeichnis manuell in einen numerischen Begriff zu transformieren ist, dann kann diese Umsetzung korrekt oder falsch erfolgen. In der manuellen Signierkontrolle kann der Fehler beseitigt werden oder unerkannt bleiben. Im Prozeß des Lochens und Lochprüfens können erneut Fehler erzeugt bzw. entdeckt werden. An die Erfassung der Daten auf maschinell lesbaren Datenträgern schließt sich ihre Verarbeitung durch die automatische Datenverarbeitungsanlage an. Unter den dort ablaufenden Eingabe-, Ausgabe-, Steuer-, arithmetischen und logischen Operationen befindet sich auch die automatische Fehlerbereinigung. Von der Fehlerstruktur der Daten und der Qualität des Programms der automatischen Kontrollen und Korrekturen hängt es ab, ob Fehler identifiziert und automatisch korrigiert werden können, oder ob sie nicht erkennbar sind und sich damit einer Korrektur entziehen. Die Übersicht zeigt weiterhin das Fehlerverhalten von Werten, die bereits bei der Erhebung falsch sind und solchen, die nicht berichtet wurden (wobei offen bleibt, ob es sich um totalen oder teilweisen Informationsausfall handelt) sowie von Werten, die geschätzt werden mußten.

Zu den manuellen Kontrollen im Vorfeld der automatischen Datenverarbeitung gehören neben dem Signieren und der Signierkontrolle Kontrollen auf Vollzähligkeit und Vollständigkeit der Belege. Wenn die Datenerfassung in Form der Beleglesung durchgeführt wird, dann sind Lesekontrollen bereits Teil der automatischen Fehlerbereinigung. In der Grafik erscheint die automatische Fehlerbehandlung als letzter Arbeitsschritt. Die Praxis zeigt jedoch, daß sich eine Phase manueller Überprüfung der Auswirkungen der maschinellen Verfahren anschließen muß.

DIE ENTSTEHUNG, IDENTIFIKATION UND BESEITIGUNG VON FEHLERN IN DEN HAUPTPHASEN DER VOLKSZÄHLUNG



Erklärung:

X = richtiger Wert

F = Fehler

L = nicht berichtet

(X) = Schätzwert

F_e = erkennbarer FehlerF_n = nicht erkennbarer Fehler
 = automatische Kontrolle
und Korrektur

Quelle: Szameitat, K., Zindler, H. J., The Reduction of Errors in Statistics by Automatic Corrections, in: Bulletin of the ISI, Proceedings of the 25th Session, 1958, p. 100.

3 Fehleridentifikation

31 Begriffsbestimmungen

Im Rahmen der Fehlerbereinigung, die alle Maßnahmen zur Qualitätsverbesserung umfaßt, wird unter Fehleridentifikation das Auffinden von ungültigen Werten und Informationsausfällen sowie von ungewöhnlichen Werten in Fragebögen, Datensätzen und Tabellen verstanden. Dieses Auffinden geschieht mit Hilfe von Kontrollen.

Kontrollen, die sich auf ein Datenfeld beziehen, heißen univariat (eindimensional, Einzelfeld-Kontrollen). Kontrollen mehrerer Datenfelder "gleichzeitig" heißen multivariat (mehrdimensional, Kombinationskontrollen). Hier ist eine Ausnahme zu beachten: Es kann gelegentlich vorkommen, daß eine Wertekombination in einem Datenfeld definiert ist.

Kontrollen, nach denen Einzelwerte oder Wertekombinationen aufgrund von Definitionen als ungültig zurückgewiesen oder als gültig angenommen werden, werden als deterministische Kontrollen (Logik-Kontrollen, Kontrollen auf der Basis theoretischen Wissens) bezeichnet. Fehler, die bei diesen Kontrollen entdeckt werden, heißen Mußfehler.

Kontrollen, nach denen Einzelwerte oder Wertekombinationen aufgrund empirischen Wissens als ungewöhnlich bezeichnet werden, sind als Wahrscheinlichkeitskontrollen definiert. Der Begriff "stochastische" Kontrollen - wie er bei Korrekturen verwendet wird - ist hier nicht üblich. Die Identifikation von Fehlern aus diesen Kontrollen werden Kannfehler genannt.

Der Begriff "deterministische" Kontrollen erscheint nicht eindeutig, da Definitionen nach der Erfahrung formuliert sein können. In vielen Fällen begründet erst die Analyse einer historischen Häufigkeitsverteilung und die Entscheidung über die Festsetzung des Annahme- bzw. Ablehnungsbereichs (Abschneidegrenzen) eine deterministische Kontrolle. In der deutschen Statistik wird der Begriff "Plausibilitätskontrolle" häufig

verwendet.¹⁾ Er bringt zum Ausdruck, daß bestimmte Werte oder Wertekombinationen logisch oder aufgrund empirischen Wissens glaubwürdig, plausibel oder konsistent sind, während andere sich ausschließen, unplausibel oder inkonsistent sind. Damit ist die Fehleridentifikation schlechthin gemeint, ohne daß zwischen deterministischen und Wahrscheinlichkeitskontrollen unterschieden wird. Der Verfasser schlägt vor, den Begriff auf Wahrscheinlichkeitskontrollen zu beschränken.

Automatische Kontrollen sind Operationen einer automatischen Datenverarbeitungsanlage zur Fehleridentifikation. Die automatische Datenverarbeitungsanlage ist ein System, das Informationen aus der Umgebung aufnimmt, speichert, verarbeitet und Informationen an die Umgebung abgibt.²⁾ Als synonyme Begriffe verwendet der Verfasser Kontrollen durch "Computer", durch "Elektronische Datenverarbeitungsanlagen (EDV)" oder "maschinelle" Kontrollen. Der automatische Ablauf der Operationen setzt die Speicherung der Daten auf Datenträger voraus.

Wenn Kontrollen mit dem Ziel der unmittelbaren Verbesserung der Daten durchgeführt werden, nennt man sie Kontrollen mit operativem Charakter. Das können Total- und Stichprobenkontrollen sein. Bei Stichprobenkontrollen tragen nur die kontrollierten und bereinigten Arbeitsmassen zur Verbesserung der Ergebnisse bei. Kontrollen, die zur nachträglichen Bewertung der Zählungsergebnisse führen, sind deskriptive Kontrollen.

Weichselberger verwendet die Begriffe Verfahrens- und Ergebniskontrollen.³⁾ Er versteht unter Verfahrenskontrollen die Überprüfung der einzelnen Zähl- und Bearbeitungsvorgänge, unter Ergebniskontrollen die Überprüfung der aggregierten Werte. Er teilt diese Kontrollen noch in endogene und exogene ein, wobei er von endogen spricht, wenn Material aus ein und derselben

-
- 1) Vgl. Szameitat, K., Möglichkeiten und Grenzen der Automatisierung in der Statistik, in: Allgemeines Statistisches Archiv, Bd 43, H.4, München 1959, S.327.
 - 2) Stichwort "Automat", Wörterbuch der Kybernetik, Hrsg. Klaus, G., Dietz, Berlin 1968, S.54.
 - 3) Weichselberger, K., Kontrollen der Ergebnisse von Volkszählungen, a.a.O., S.14.

Zählung zu Kontrollzwecken herangezogen wird, während bei ihm exogene Kontrollen auf den Informationen anderer Statistiken beruhen.

Während den bisher genannten Kontrollen gemeinsam ist, daß sie die Ausprägungen von Merkmalen der statistischen Einheiten behandeln, besteht die Aufgabe der Bestandskontrollen darin, die regionale Zuordnung der statistischen Einheiten zu sichern. Bestandskontrollen werden als automatische Kontrollen angesehen. Vollzähligkeitskontrollen dagegen, d.h. Feststellungen, ob die Fragebögen aus den Gemeinden vollzählig zur Aufbereitung gelangt sind, werden den manuellen Arbeiten im Vorfeld der automatischen Datenverarbeitung zugerechnet.

32 Wahrscheinlichkeitskontrollen

In einem Datenbestand können Werte ungewöhnlich sein, weil sie Fehler enthalten. Werte können aber auch ungewöhnlich und dennoch richtig sein. Für Kontrollzwecke ist es notwendig, ungewöhnlichen Werten ein Wahrscheinlichkeitsmaß zuzuordnen. Dazu ist es im Prinzip notwendig, vor Beginn der Kontrollen die Häufigkeitsverteilungen der zu kontrollierenden Variablen zu kennen. Da dies unmöglich ist, müssen Annahmen über die zugrunde liegende Verteilung der Variablen gemacht werden. Es wird also die Frage nach der Wahrscheinlichkeit gestellt, daß ein beobachteter Wert von dem "wahren" Wert einer historischen Verteilung oder einer Bezugsgröße aus der aktuellen Verteilung, der er entstammt (repräsentiert z.B. durch eine Entscheidungss Stichprobe), um einen bestimmten Betrag abweicht. Ist die Differenz groß, gilt der beobachtete Wert als Extremwert oder "Ausreißer". Ein beträchtlicher Teil der Ausreißertheorie befaßt sich mit dem univariaten parametrischen Fall unter der Annahme, daß die untersuchten Variablen als etwa normalverteilt angesehen werden können. Auch komplexere Tatbestände, die bei der Kombination von Variablen (multivariater Fall) gegeben sind, lassen sich mit Hilfe der Normalitätsannahme noch recht gut lösen. Wenn diese Prämisse als zu einengend aufgegeben wird, lassen sich Gruppen von Verteilungen mit spezifischen Streuungsverhältnissen entwickeln, wobei die Anpassung der zu kontrollierenden Variablen an die Struktur mit der größten Übereinstimmung vorgenommen wird. Diese Kontrollen

wie auch Verfahren auf der Grundlage des Bayes'schen Theorems werden unter den "sonstigen parametrischen Verfahren" behandelt. Es folgen Kontrollen, die unabhängig sind von einer Annahme über die Gestalt der zugrunde liegenden Verteilung und die als nichtparametrische Tests bezeichnet werden.

321 Parametrische Verfahren

3211 Univariate Verfahren

Als Vergleichsgrundlage für die zu kontrollierenden Werte ist eine historische Verteilung ideal, die der aktuellen Situation angepaßt ist. Da solche Angleichungen nur im Falle kurzfristig wiederholter Erhebungen brauchbare Lösungen ergeben, gilt es also, nach Ersatzlösungen zu suchen. Wenn eine Normalverteilung unterstellt werden kann, bei der Mittelwert und Standardabweichung im Zeitablauf unverändert geblieben sind, dann ist folgendes Vorgehen für eine einzelne Variable möglich: Differenzbildung zwischen den Beobachtungswerten der aktuellen Zählung und dem Mittelwert der historischen Grundgesamtheit, Division durch die Standardabweichung der historischen Grundgesamtheit und Bestimmung des "Maßes an Unsicherheit" für das Auftreten einzelner Beobachtungswerte mittels der Standardnormalverteilung.

Wenn jedoch vermutet werden muß, daß Mittelwert und/oder Standardabweichung sich im Zeitlauf verändert haben, dann müssen diese Parameter aus der aktuellen Zählung berechnet werden. Wenn eine nicht allzu große Stichprobe gezogen wird zur Schätzung der Parameter, sind die Zufallsschwankungen zu berücksichtigen. Das größere Problem aber ist, daß Mittelwert und Standardabweichung auf einer Stichprobe beruhen, die selbst fehlerhaft ist. Es wird also ein Ausreißer durch seine besonders große Abweichung zum Mittelwert identifiziert, wobei der Mittelwert unter Einschluß dieses Ausreißers zustande gekommen ist. Es gibt verschiedene Versuche, das Problem zu lösen.

32111 Gross Error-Methode

Gegeben sei eine normalverteilte Grundgesamtheit $N(\mu, \sigma^2)$. In

Stichproben aus dieser Gesamtheit seien verfälschte Beobachtungen enthalten. Angenommen, die Verzerrungen ließen sich wie folgt typisieren:

1. Beobachtungen aus $N(\mu + \lambda\sigma, \sigma^2)$
2. Beobachtungen aus $N(\mu, \lambda^2\sigma^2)$

Es wird unterstellt, daß viele Fehler eine Kombination beider Typen darstellen. Dafür ist der Begriff "gross error" geprägt worden.¹⁾

Die Identifikation und Rückweisung solcher Fehler soll ebenfalls mittels statistischer Tests erfolgen. Dixons Lösungsvorschlag basiert auf der Kenntnis der Verteilung von Quotienten der Form

$$r = \frac{x_n - x_{n-j}}{x_n - x_i}$$

für sehr kleine Werte von i und j sowie Stichprobenumfängen von $n = 3, \dots, 30$. Die untersuchten Variablen stammen aus geordneten Stichproben, d.h. für die Beobachtungswerte gilt $x_1 \leq x_2 \leq \dots \leq x_n$. Dixon hat für eine Auswahl von Quotienten unter Berücksichtigung unterschiedlicher Signifikanzniveaus Wahrscheinlichkeiten tabelliert.²⁾ Die Effizienz des Verfahrens wird von Dixon detailliert an jedem der beiden genannten Fehlertypen untersucht.

Als Beispiel für die Vorgehensweise soll ein bestimmter Quotient betrachtet werden:

$$r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1} ; \quad n = 3$$

Drei aufeinander folgende Beobachtungen seien der Größe nach in aufsteigender Reihenfolge sortiert. Es werden folgende Differenzen gebildet:

$$\begin{aligned} x_3 - x_1 &= a \\ x_3 - x_2 &= b \\ x_2 - x_1 &= c \end{aligned}$$

Testgrößen sind b/a und c/a . Diese werden - bei einem Signifikanzniveau von $\alpha = 5\%$ - mit dem Tabellenwert 0,941 verglichen. Überschreitet eine Testgröße diesen Wert, dann ist x_3 bzw. x_1 als ungewöhnlich zu kennzeichnen.

1) Dixon, W. J., Analysis of Extreme Values, in: The Annals of Mathematical Statistics, Vol. 21, Waverly Press, Baltimore 1950, S.489.

2) Dixon, W. J., Ratios Involving Extreme Values, in: The Annals of Mathematical Statistics, Vol. 22, Waverly Press, Baltimore 1951, S.73 ff.

Das Maß für die Ungewöhnlichkeit der Quotienten ist abhängig von der Variabilität der Beobachtungen innerhalb selbst kleinster Gruppen ($n \geq 3$), für die eine Normalverteilung unterstellt wird. Anders als bei vielen anderen Kontrollverfahren werden hier gleichzeitig mehrere Beobachtungen behandelt.

Modifikationen des Verfahrens bestehen in der Wahl der Gruppengröße und damit in der Gestaltung der Relation zwischen Prüfaufwand und Prüferfolg. Außerdem können Beobachtungen der jeweils vorhergehenden Gruppe in die nächste einbezogen werden, um so 2 oder 3 hintereinander liegende Werte als extrem kennzeichnen zu können.

Erfahrungen mit der Methode zeigen, daß die Beobachtungen stark verzerrt sein müssen, damit sie identifiziert werden können.¹⁾

32112 FUNOR (Full Normal Rejection)

FUNOR ist ein von Tukey entwickeltes Verfahren der Fehleridentifikation.²⁾ Es setzt eine normalverteilte Grundgesamtheit voraus. Wenn Mittelwert und Standardabweichung aus der zu kontrollierenden Stichprobe entwickelt werden müssen, dann sind sie durch etwa vorhandene Ausreißer verzerrt. Da beide Parameter auf Extremwerte besonders stark reagieren, schlägt Tukey vor, die Einzelbeobachtungen auf den Median zu beziehen. Alternativ könnte auch der Mittelwert einer "getrimmten" Reihe Verwendung finden, d.h. einer Reihe geordneter Beobachtungen, von der eine gewisse gleich große Anzahl kleinster sowie größter Werte abgeschnitten und in die Mittelwert-Berechnung nicht einbezogen wird. Bei FUNOR wird nur das obere und untere Drittel der geordneten Stichprobenwerte untersucht.

1) Vgl. Naus, J. I., Data Quality Control and Editing, Statistics: Textbooks and Monographs, Vol. 10, Marcel Dekker, New York 1975, S.43.

2) Tukey, J. W., The Future of Data Analysis, in: The Annals of Mathematical Statistics, Vol. 33, Waverly Press, Baltimore 1962, S.22 ff.

Berechnet werden:

$$z_i = \frac{x_i - \tilde{x}}{a_{i/n}} ;$$

x_i : geordnete Beobachtungen
 $\tilde{x}$: Median
 i : Rangziffer

Als $a_{i/n}$ wird der "typische" Wert für die i -te unter n geordneten Beobachtungen einer Standardnormalverteilung bezeichnet. Erfahrungsgemäß gilt

$$\Phi(a_{i/n}) = (3i - 1)/(3n + 1); \Phi(a) \triangleq \text{Standardnormalverteilung}$$

als gute Näherung.¹⁾ Nach Berechnung des Quotienten mittels Rangziffer und Stichprobenumfang kann $a_{i/n}$ aus der Standardnormaltabelle abgelesen werden. Im unteren Drittel der Verteilung wird wegen der Symmetrie $-a_{1/n} = a_{n/n}$ gesetzt usw. So ist z.B. $-a_{1/14} = a_{14/14} = \Phi(2/43) = 1,685$.

Schließlich wird der Median $\tilde{z}$ aus den z_i -Werten bestimmt.

"... $\tilde{z}$ ist eine vernünftige Schätzung - mit leichter Tendenz zur Übertreibung - von σ des 'vorwiegend normalen Teils' der Verteilung, aus der die x_i stammen."²⁾

Es werden zwei Kontrollen durchgeführt:

$$|x_i - \tilde{x}| \geq A \cdot \tilde{z} \quad (1. \text{ Vergleich})$$

$$z_i \geq B \cdot \tilde{z} \quad (2. \text{ Vergleich})$$

Die Empfindlichkeit des Verfahrens kann über die - subjektiven - Vorgaben von A und B gesteuert werden. Je größer A und B sind, umso ungewöhnlicher müssen die beobachteten Werte sein, um als extrem gekennzeichnet zu werden. Ein Wert wird zurückgewiesen, wenn er beide Schranken erreicht bzw. übertrifft.

Eine Eigenart der Methode besteht darin, daß Werte, die z.B. an k -ter und $(k+2)$ -ter Stelle in der Rangfolge liegen, als ungewöhnlich beanstandet werden, der $(k+1)$ -te jedoch nicht. Das liegt daran, daß das Verfahren für die Beobachtungswerte strikt eine Normalverteilung voraussetzt und die Einzelwerte an der Normalstruktur mißt. Das Erfordernis der Normalität ist der Haupteinwand gegen das Verfahren. Dennoch wird es als sehr empfindlich und flexibel beurteilt.³⁾

1) Tukey, J. W., The Future of Data Analysis, a.a.O., S.22.

2) Ders., ebenda, S.23.

3) Vgl. Naus, J. I., Data Quality Control and Editing, a.a.O., S.49.

32113 Annahme-Kontrolle

Dieses Verfahren hat Ähnlichkeit mit den Annahmestichproben in der Qualitätskontrolle. Während diese aber im Vorfeld der automatischen Fehlerbereinigung (manuelle Signierkontrolle, Prozeß des Lochprüfens) verwendet werden, ist die Annahme-Kontrolle als automatische Kontrolle konzipiert. Während in der Statistischen Qualitätskontrolle der α -Fehler (Produzentenrisiko) möglichst klein gehalten werden soll, kommt es hier stärker auf den β -Fehler (Konsumentenrisiko) an, d.h. das Risiko soll klein gehalten werden, "schlechte" Werte anzunehmen. Voraussetzung für dieses Verfahren ist wieder, daß die zu kontrollierenden Variablen normalverteilt sind.

Als sinnvolle Kontrollvariable wird hier die Standardabweichung der zu kontrollierenden Variablen angesehen. Nach einer in der Praxis im Sinne des Veröffentlichungsprogrammes im allgemeinen durchgeführten Gruppierung des zu prüfenden Materials, wird für jede Gruppe nach subjektivem Ermessen eine Obergrenze für die Standardabweichung vorgegeben:

$$s \leq G_0 \quad (G_0: \text{Obergrenze})$$

Wie bei Qualitätskontrollen üblich, werden nicht alle Einzelwerte kontrolliert, sondern aus jeder Gruppe wird eine Zufallsstichprobe gezogen. Der Stichprobenumfang n je Gruppe kann wieder über Operationscharakteristiken bestimmt werden, wenn Produzentenrisiko, Konsumentenrisiko, Gut- und Schlechtgrenze bekannt sind. Wenn α , β und das Verhältnis von Gut- zu Schlechtgrenze gruppenweise nicht voneinander abweichen, dann ist n jeweils gleich groß und unabhängig von der Größe der Gruppen.

Aus der Stichprobenvarianz s^2 wird die Testgröße

$$k^2 = (n - 1) \cdot s^2 / G_0^2$$

berechnet, die wegen der vorausgesetzten Normalverteilung χ^2 -verteilt ist. Übersteigt dieser Wert k^2 das entsprechend der Risikobereitschaft gewählte χ^2 -Percentil, dann wird die gesamte Gruppe als fehlerhaft zurückgewiesen, sonst wird sie angenommen. Natürlich besteht das Risiko, daß s nicht im Annahmebereich liegt.

Die Verwendung dieses Kontrollverfahrens setzt voraus, daß die Gefahr, Verfälschungen in Untergruppen infolge nicht geprüfter Materialteile zu erhalten, als nicht zu hoch eingeschätzt wird. Die Ergebnisse sind selbstverständlich nicht unabhängig von einer sinnvollen Gruppenbildung.

3212 Multivariate Verfahren

32121 Bivariater Fall

Es kann vorkommen, daß ein Wert von einer Einzelfeld-Kontrolle nicht beanstandet wird, während er in einer Wertekombination als ungewöhnlich gekennzeichnet wird. Andererseits könnte er sich bei einer Regressionsanalyse als ein zu vernachlässigender Residualwert erweisen. Der Zusammenhang soll am Beispiel der Merkmalskombination "Körpergewicht" und "Körpergröße" erläutert werden. Auch hier läßt sich wieder unterscheiden zwischen einer Normalverteilung, deren Parameter im Zeitablauf unverändert geblieben sind, so daß historische Werte als Vergleichsgrundlage für die zu kontrollierenden Werte dienen können, und einer Normalverteilung, deren Parameter aus einer Stichprobe der aktuellen Werte geschätzt wurden.

Der Quotient

$$\frac{\text{beobachtetes Gewicht} - \text{Durchschnittsgewicht für diese Größe}}{\text{Standardabweichung der Gewichte bei dieser Größe}}$$

wird auch gelegentlich als "standardisierter Residualwert" bezeichnet.¹⁾ Wenn Durchschnittsgewicht und Standardabweichung der Gewichte bei gegebener Größe für eine statistische Einheit bekannt sind, dann lassen sich die Wahrscheinlichkeitsmaße für den Quotienten wieder der Standardnormaltabelle entnehmen.

Von ganz entscheidender Bedeutung für die Fehlerbeseitigung ist, daß bei Rückweisung des Quotienten ohne zusätzliche Informationen die für die Beanstandung verantwortliche Variable

1) Naus, J. I., Data Quality Control and Editing, a.a.O., S.51.

nicht bestimmt werden kann. Die Problematik wird im Kapitel "Fehlerallokation" behandelt. Wenn die zu Kontrollzwecken benötigten Werte nicht zur Verfügung stehen, kann versucht werden, sie mit Hilfe eines Modells (z.B. der linearen Regression) abzuleiten. Im vorliegenden Beispiel kann für jede Körpergröße das Durchschnittsgewicht bestimmt werden. Wenn anzunehmen ist, daß die Standardabweichungen der Gewichte für unterschiedliche Größen dieselben sind, dann läßt sich die "allgemeine" Standardabweichung als "Standardfehler der Schätzung der Regressionskurve" berechnen. Der standardisierte Residualwert lautet dann:

$$\frac{\text{beobachtetes Gewicht} - \text{vom Modell bestimmtes Durchschnittsgewicht für diese Größe}}{\text{Standardfehler der Schätzung der Regressionskurve}}$$

Iterative Verfahren können gute Dienste bei der Eliminierung von Ausreißern leisten. Naus weist auf ein Maschinenprogramm von Scott hin,¹⁾ bei dem die Regressionsschätzung im ersten Schritt auf alle Daten angewandt wird, um die größten Ausreißer zu beseitigen. Im nächsten Schritt wird das Modell auf die im ersten Schritt bereinigten Daten angewandt und die nächst größeren Ausreißer aussortiert. Diese Prozedur kann fortgesetzt werden. Wenn es keine extremen Werte gibt, bleibt das Regressionsmodell von einer Iteration zur nächsten stabil. Ansonsten stabilisiert sich das Modell bereits nach zwei oder höchstens drei Stufen.

Das Beispiel der Verknüpfung von zwei Variablen kann verallgemeinert werden auf den multivariaten Fall. Dann wird mit Hilfe des Regressionsmodells eine Variable aus den übrigen geschätzt, und die Werte dieser Variablen werden mit den beobachteten verglichen.

In USA wurden bei der Handwerkszählung 1963 Erfahrungen mit dem linearen Regressionsmodell gemacht. Dabei waren rd. 24 000 Regressionen geschätzt worden. Die Ergebnisse waren nicht befriedigend. Man hielt einfachere Quotienten-Kontrollen für effizienter.²⁾

1) Naus, J. I., Data Quality Control and Editing, a.a.O., S.52.

2) Ders., ebenda, S.53.

32122 Hauptkomponentenlösung

Ein Datensatz, bestehend aus Datenfeldern, wird als Vektor dargestellt:

$$\underline{x} = (x_1, x_2, \dots, x_n)$$

Zu diesem beobachteten Datensatz soll ein entsprechender fehlerfreier Vektor $\underline{y}$ existieren. Der Fehlervektor ist $\underline{e}$. Das Modell hat die Form:

$$\underline{x} = \underline{y} + \underline{e}$$

Es wird weiter angenommen, daß $\underline{y}$ normalverteilt $N(\xi, \Sigma)$ und ξ und Σ bekannt sind. Die y_i seien standardisiert. Die Kontrolle auf Extremwerte ist definiert als Test der Nullhypothese $H_0: \underline{e} = 0$ gegen die Alternativhypothese $H_1: \underline{e} \neq 0$. Da sich die Fehleridentifikation auf einzelne Komponenten von $\underline{x}$ beziehen soll, wird eine eingeschränkte Alternativhypothese definiert:

$$\begin{aligned} H_2: e_j &\neq 0 \quad (\text{für ein unbekanntes } j) \\ e_i &= 0 \quad (\text{für alle } i \neq j) \end{aligned}$$

Hawkins untersucht fünf aus der Ausreißer-Theorie bekannte Teststatistiken auf ihr Verhalten.¹⁾

$$T_0 = \max (|x_1|, |x_2|, \dots, |x_n|)$$

$$T_1 = \underline{x}' \Sigma^{-1} \underline{x}$$

T_1 ist ein Likelihood-Quotienten-Test. T_0 ist mächtiger als T_1 gegen die Alternativhypothese H_2 , wenn Σ diagonal ist. Da Σ im allgemeinen nicht diagonal ist, sollte $\underline{x}$ in einen Vektor $\underline{z}$ transformiert werden, dessen Kovarianzmatrix diagonal ist, um dann T_0 auf die transformierten Variablen anzuwenden.

Wenn $A = (a_{ij})$ die Matrix der Eigenvektoren von Σ ist und Λ die Diagonalmatrix der Eigenwerte λ_i mit $(\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k)$, dann ist

$$\underline{w} = A \underline{x}$$

normalverteilt $N(0, \Lambda)$ und die w_i sind unabhängige Variable mit Varianz λ_i .

1) Hawkins, D. M., The Detection of Errors in Multivariate Data Using Principal Components, in: Journal of the American Statistical Association, Vol. 69, No. 346, Menasha, Wisc. 1974, S.340-344; Fellegi, I. P., Automatic Editing and Imputation of Quantitative Data, hekt. Arbeitsexemplar, Toronto 1975; zur Ausreißer-Theorie vgl. Gumbel, E. J., Statistics of Extremes, Columbia University, Fourth Printing, New York 1967.

Die Transformation

$$\underline{z} = B \underline{x}$$

($b_{ij} = a_{ij} / \sqrt{\lambda_i}$ führt zu Variablen z_i , die normal und unabhängig sind sowie Varianz 1 haben. Hawkins folgert nun, daß $\sum b_{ij} e_j$ der Erwartungswert von z_i unter H_2 ist. H_0 kann nun gegen H_2 getestet werden, indem z_i auf Mittelwert 0 getestet wird. Die Mächtigkeit des Tests hängt davon ab, wie groß die b_{ij} sind. Da

$$\sum b_{ij}^2 = 1/\lambda_i$$

folgt, daß die z_i , die großen λ_i entsprechen, unempfindlich gegen Fehler sind und daher wegfallen sollten. Auf diese Weise verbessert sich die Mächtigkeit der beiden Tests oben bei gegebenem Signifikanzniveau. Wenn n_0 der λ_i als "klein" nach bestimmten Kriterien ausgewählt werden können, dann können nach Hawkins zwei weitere Tests definiert werden:

$$T_2 = \sum_{i=1}^{n_0} z_i^2 \quad \text{und}$$

$$T_3 = \max_{1 \leq i \leq n_0} |z_i|$$

Für $n_0 = n$ sind die Tests T_2 und T_1 identisch.

Die letzte von Hawkins vorgeschlagene Teststatistik ist eine Variante von T_3 . Solange die Variablen z_i unkorreliert sind, Mittelwert 0 und Varianz 1 haben, erhält jede orthogonale Transformation diese Eigenschaften aufrecht. Nach orthogonaler Transformation C der z erhält er $\underline{v}$:

$$\underline{v} = C\underline{z} = CB\underline{x} = D\underline{x}$$

Wenn C geeignet gewählt wird, kann jede Spalte von D durch ein einziges zahlenmäßig großes Element beherrscht werden, wobei der Rest nahe null ist. Auf diese Weise beeinflusst ein Fehler in einer einzigen Komponente von $\underline{x}$ primär eine Komponente von $\underline{z}$. Nach den von Dixon erzielten Ergebnissen bringt

eine Variante von T_3 bezüglich der v bessere Ergebnisse. D kann mittels Varimax-Rotation der ersten n_0 Zeilen von B erhalten werden, während die übrigen unverändert bleiben sollen.

Die Variante von T_3 lautet:

$$T_4 = \max_{1 \leq i \leq n_0} |v_i|$$

Fellegi stellt fest,¹⁾ daß T_0 bis T_2 immer gelten, unabhängig davon, ob y normalverteilt ist oder nicht. Für T_3 und T_4 können kritische Werte nur dann aus der Normalverteilung entnommen werden, wenn z und y nicht nur unkorreliert sondern auch unabhängig sind, und das ist nicht notwendigerweise so, wenn y nichtnormal ist. Nichtnormale Verteilungen können möglicherweise entsprechend transformiert werden.

Von noch größerer Bedeutung sind die Ränder der Häufigkeitsverteilungen. Mit Hilfe einer sorgfältig fehlerbereinigten Stichprobe können ξ und Σ geschätzt und die Gültigkeit der Formel oben kontrolliert werden. Damit läßt sich unter Umständen auch eine geeignete Transformation der empirischen Verteilung durchführen. Auch durch Schichtenbildung läßt sich die Variabilitätsstruktur verbessern. Die Abschneidegrenzen können aber auch empirisch abgeleitet werden anstatt Normalität zu unterstellen. Allerdings geht es bei der Fehlerbereinigung, die das Ziel hat, Verzerrungen in den Verteilungen der kontrollierten und korrigierten Datensätze zu minimieren, nicht so sehr darum, die exakte Größe des kritischen Bereichs zu kennen, als vielmehr darum, die Auswirkungen unterschiedlicher Abschneidegrenzen zu untersuchen.

1) Fellegi, I. P., Automatic Editing and Imputation of Quantitative Data, a.a.O., S.5.

In Simulationsläufen fand Fellegi heraus,¹⁾ daß 74 % der Datensätze mit einem Fehler von 0,5 Standardabweichungen in einem der Felder entdeckt wurden, 97,5 % mit einem Fehler von 1 Standardabweichung und 100 % mit einem Fehler von 1,5 Standardabweichungen. Die Mächtigkeit der Tests hängt entscheidend von der multiplen Korrelation der Datenfelder ab. Wenn diese hoch ist, ist die Fehleridentifikation auf multivariater Grundlage effektiver als auf univariater. T_2 und T_4 erwiesen sich als wirksamer gegen Einzelfehler in einzelnen Komponenten als der Standard- χ^2 -Test. Zwar ist T_2 weniger mächtig als T_4 gegen Einzelfehler, seine Verwendung ist jedoch anzuraten bei Kombinationsfehlern. Bei Wahl der geeigneten Transformationsmatrix ist T_4 gegenüber T_3 überlegen. Die Testergebnisse verbessern sich deutlich, wenn die Hauptkomponenten mit den höchsten Varianzen entfernt werden.

Diese Lösung bringt auch Vorteile bei der Quotienten-Kontrolle.²⁾ Bisher war es so, daß bestimmte Kontrollgrenzen vorformuliert wurden, z.B.

$$20 \leq x_1/x_2 \leq 30$$

In linearer Form entspricht das den Gleichungen

$$\begin{aligned} 20x_2 - x_1 &\leq 0 \text{ und} \\ x_1 - 30x_2 &\leq 0 \end{aligned}$$

Rückweisungen erfolgten bei Unter- oder Überschreiten dieser Grenzen. Fellegi weist nach, daß standardisierte Hauptkomponenten vorformulierte Ungleichheitsbedingungen wirksam ersetzen können. Der größte Erfolg bei der Identifikation von Ausreißern wird durch lineare Transformationen von $\underline{x}$ mit der kleinsten Varianz erzielt. Hauptkomponenten liefern unter allen linearen Transformationen einige mit den vergleichsweise kleinsten Varianzen. Voraussetzung ist, daß jede Hauptkomponente zur vorhergehenden unkorreliert ist. Die Hauptkomponenten führen, wenn sie unkorreliert sind und so für praktische Zwecke vermutlich

1) Fellegi, I. P., Automatic Editing and Imputation of Quantitative Data (Summary), Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975.

2) Ders., Automatic Editing and Imputation of Quantitative Data, a.a.O., S.16 ff.

fast unabhängig, selbst zur Berechnung geeigneter Vertrauensgrenzen und stellen auf diese Weise sicher, daß nicht mehr als ein bestimmter Anteil fehlerfreier Datensätze zurückgewiesen wird. In einem Theorem wird unter allgemeinen Bedingungen nachgewiesen, daß strukturelle Beziehungen zwischen den Datenfeldern der Datensätze sich zwischen den Hauptkomponenten widerspiegeln. Wenn jedoch der Inhalt eines Datenfeldes sich aus den Inhalten einiger anderer Datenfelder durch Summierung ergibt, also exakte Identität hergestellt werden muß, ist die Vorformulierung von Kontrollgrenzen unerlässlich.

322 Sonstige parametrische Verfahren

3221 Anpassung an das Elderton-Pearson'sche System von Häufigkeitsverteilungen

Lösungen, die die Annahme der Normalität voraussetzen, führen zwar zu effizienten Verfahren der Fehleridentifikation, sind jedoch meist realitätsfern. O'Reagan entwickelt eine Lösung für Wahrscheinlichkeitskontrollen,¹⁾ bei der eine Reihe von Variabilitätsstrukturen verschiedenen Typen (parametrischen Verteilungsklassen) zugeordnet werden. Die Verteilungsformen jedes Typs sind durch eine Kombination von Parametern definiert. Es sind unterschiedliche Klassifikationsschemata entwickelt worden, z.B. von Edgeworth und Charlier und von Elderton-Pearson. Das Kriterium für die Anpassung muß sein, gute Lösungen für ungewöhnliche Werte (also an den Rändern der Verteilungen) zu erzielen. Oftmals gelingt dagegen lediglich die Anpassung für die häufig auftretenden Werte gut.

O'Reagan wählt das Klassifikationsschema von Elderton-Pearson, wonach die zu kontrollierende Verteilung danach untersucht wird, ob sie näherungsweise mit einer von 13 Häufigkeitsverteilungen übereinstimmt. Die Anpassung erfolgt innerhalb jedes Typs durch Schätzung von vier Parametern: Mittelwert, Standardabweichung und zwei höhere Momente.

Die Vorteile bei der Anwendung dieser Technik bestehen darin,

1) O'Reagan, R. T., Possible Techniques for Computer Editing of Data, US Bureau of the Census, Washington, D.C. 1965.

daß - obwohl die Normalitätsannahme fallengelassen wird - dennoch eine kompakte Beschreibung der Variabilitätsstruktur gegeben wird. Das ist ein Vorteil gegenüber verteilungsfreien Lösungen.

Der Nachteil besteht darin, daß die Kontrollsituation nur schwer in Maschinenprogramme übersetzt werden kann, und die Ausführung durch die EDV kompliziert ist.

3222 Kontrolle auf der Grundlage des Bayes'schen Theorems

In diesem Lösungsvorschlag¹⁾ wird keine bestimmte Annahme über die Parameter der zu kontrollierenden Gesamtheit gemacht. Jede statistische Einheit in einer Erhebung (das entspricht wieder einem Datensatz) wird als p-dimensionaler Vektor angesehen. Es geht hier nicht um die Entscheidung über Annahme oder Rückweisung einzelner Werte dieses Vektors sondern der gesamten Einheit. Das Problem der Fehlerallokation wird ausgeklammert.

Die Autoren nutzen den Formalismus des Bayes'schen Theorems für die Berechnung der Wahrscheinlichkeiten für die richtige bzw. falsche Entscheidung darüber, ob eine statistische Einheit fehlerfrei oder fehlerhaft ist. Wenn $x_j \in A$, dann ist x_j fehlerfrei; wenn $x_j \in \bar{A}$, dann ist x_j fehlerhaft. Richtige bzw. falsche Entscheidungen über die Zugehörigkeit der x_j , die relativ leicht zu treffen sein sollen, werden durch die "Entscheidungsmengen" B bzw. $\bar{B}$ angegeben. Es gelten die Entscheidungsregeln:

$$\begin{aligned} x_j \in B &\rightarrow x_j \in A \text{ und} \\ x_j \in \bar{B} &\rightarrow x_j \in \bar{A} \end{aligned}$$

Das Modell beruht auf der Kenntnis der Wahrscheinlichkeiten dafür, daß die statistischen Einheiten zu fehlerfreien bzw. fehlerhaften Mengen A bzw. $\bar{A}$ gehören. Außerdem müssen die Wahrscheinlichkeiten für richtige bzw. falsche Entscheidungen (B bzw. $\bar{B}$) gegeben sein.

1) Cassel, P. G., Fisz, A., Statistical Problems in Automatic Control of Data, in: Statistisk Tidskrift, Hrsg. Statistiska Centralbyrån, H. 2, Stockholm 1970, S.96-99.

Symbole für diese Wahrscheinlichkeiten seien:

$$P(A) = \pi_1; \quad P(\bar{A}) = \pi_2 = 1 - \pi_1$$

$$P(B|A) = p_1; \quad P(B|\bar{A}) = p_2$$

$$P(\bar{B}|A) = q_1 = 1 - p_1; \quad P(\bar{B}|\bar{A}) = q_2 = 1 - p_2$$

π_1 sei z.B. die Wahrscheinlichkeit, daß x_j zur fehlerfreien Menge A gehört. p_1 sei z.B. die Wahrscheinlichkeit für die richtige Entscheidung ("Entscheidungs Menge" B) unter der Bedingung, daß x_j zur fehlerfreien Menge A gehört.

Nach dem Bayes'schen Theorem lauten die a posteriori-Wahrscheinlichkeiten für richtige Entscheidungen:

$$P(A|B) = \frac{\pi_1 p_1}{\pi_1 p_1 + \pi_2 p_2}$$

$$P(\bar{A}|\bar{B}) = \frac{\pi_2 q_2}{\pi_1 q_1 + \pi_2 q_2}$$

Diese Wahrscheinlichkeiten sollten möglichst groß sein.

Die a posteriori-Wahrscheinlichkeiten für falsche Entscheidungen lauten:

$$P(\bar{A}|B) = \frac{\pi_2 p_2}{\pi_1 p_1 + \pi_2 p_2}$$

$$P(A|\bar{B}) = \frac{\pi_1 q_1}{\pi_1 q_1 + \pi_2 q_2}$$

Diese Wahrscheinlichkeiten sollten möglichst klein sein.

Um B praktisch bestimmen zu können, muß die Wahrscheinlichkeitsstruktur von A und $\bar{A}$ bekannt sein. Am univariaten Fall wird erläutert, wie durch Vergleich einer praktisch durchgeführten Stichprobenerhebung mit einer fehlerfreien Kontrollerhebung desselben Umfangs Schätzwerte für $P(A)$ und $P(\bar{A})$ gewonnen werden. Die Autoren machen keine konkreten Aussagen über die Häufigkeitsverteilungen $f(x|A)$ und $f(x|\bar{A})$ und somit über die Größe des Intervalls B unter diesen Kurven. Vorausgesetzt, die beiden Funktionen seien bekannt, so wird vorgeschlagen, das Bayes'sche Theorem wiederholt anzuwenden und dabei B zu variieren, um numerisch zu einer optimalen Lösung zu gelangen. Allerdings ist zu berücksichtigen, daß zwei Wahrscheinlichkeiten für richtige Entscheidungen zu maximieren sind, d.h. die Lösung erfordert ein Urteil über die Gewichte, die beiden Wahrscheinlichkeiten zuzuordnen sind.

323 Nichtparametrische Verfahren

Nichtparametrische Verfahren sind dadurch gekennzeichnet, daß sie ohne Verteilungsannahmen auskommen. Es genügt die Forderung großer Ähnlichkeit zur historischen Verteilung. Zwei nichtparametrische Lösungen zur Fehleridentifikation werden hier dargestellt. Bei der ersten wird unterstellt, daß die Spannweite historischer "wahrer" Werte geeignet ist zur Kontrolle aktueller Werte. Bei der zweiten werden im Prinzip nur Informationen über Mittelwert und Varianz benötigt.

3231 Order Statistics-Lösung

Die Anwendung dieses Verfahrens setzt die Anordnung der beobachteten Werte in einer Ausgangsstichprobe (initial sample) voraus. Aus der Literatur¹⁾ ist nicht zu entnehmen, ob die Ausgangsstichprobe aus einer historischen Verteilung oder der aktuellen gezogen wird. Wenn der kritische Bereich aus einer historischen Verteilung noch gültig ist, dann ist ein verlässlicher Kontrollmaßstab gegeben. Eine Stichprobe aus aktuellen Werten müßte vermutlich erst gesichert werden.

Aus N Werten wird die Spannweite ermittelt. Wenn die zu kontrollierenden Werte in zufälliger Reihenfolge auftreten, dann nimmt die Wahrscheinlichkeit, daß die $(N + 1)$ te Beobachtung zufällig die größte ist, den Wert $1/(N + 1)$ an. Entsprechendes gilt für die kleinste Beobachtung. Die Wahrscheinlichkeit, daß die $(N + 1)$ te Beobachtung zufällig außerhalb dieser Grenzen liegt, beträgt $2/(N + 1)$.

Unter Abwägung von Genauigkeits- und Wirtschaftlichkeitsansprüchen sollte der Umfang der Ausgangsstichprobe zwischen 10 und 100 Einheiten betragen. Nach Anordnung der Werte wird der kleinste Wert als Untergrenze und der größte Wert als Obergrenze bestimmt. Für die $(N + 1)$ te Beobachtung wird festgestellt, ob sie sich innerhalb der Spannweite befindet. Wenn das der Fall

1) Vgl. Naus, J. I., Data Quality Control and Editing, a.a.O., S.55-57.

ist, bleiben die Grenzen unverändert und der Nenner erhöht sich um 1. Wenn ein Wert nicht innerhalb der Spannweite liegt, wird der Quotient $2/(N + x)$ ausgerechnet. Falls die Beobachtung überprüft werden kann und sich herausstellt, daß sie zulässig ist, dann werden die Grenzen und der Stichprobenumfang verändert. Für den Fall, daß die Beobachtung korrigiert werden muß und dann in den Toleranzbereich fällt, bleiben die Grenzen unverändert und der Stichprobenumfang erhöht sich um 1.

Strenge Voraussetzung für das Verfahren ist die zufällige Reihenfolge der Beobachtungen. Wenn die Daten gruppiert sind, dann verlangt jede Schicht ihre eigenen Grenzen. Wichtig ist, daß sich die Grenzen auf für richtig befundene Werte beziehen, so daß als Mindestforderung die Anfangs- und Endwerte der Spannweite exakt zu überprüfen sind. Bei Verwendung großer Datenmengen nähert sich der Anteil richtiger Werte, der zurückgewiesen wird, dem Wert 0. Um die Grenzen nicht zu weit zu ziehen, sollte versucht werden, das historische Wissen auf kleinere Stichproben zu beschränken, um die Effizienz des Verfahrens nicht zu gefährden. Allerdings lassen sich bei größeren Stichproben statt des höchsten und des niedrigsten Wertes auch andere Grenzen formulieren, damit die Wahrscheinlichkeit, Werte zu entdecken, nicht zu gering wird. Es lassen sich auch Werte für den Fall entwickeln, daß mehrere Beobachtungen der zu prüfenden Zählung gleichzeitig außerhalb der formulierten Grenzen liegen. Größte Schwierigkeiten dürfte es in der Praxis machen, die Gültigkeit des Annahme- bzw. Ablehnungsbereiches sicherzustellen und die Zufallsfolge der zu prüfenden Werte zu gewährleisten.

3232 Ungleichungen vom Tschebyscheff-Typ

Bei diesen Verfahren wird angenommen, daß die zu kontrollierenden Werte aus einer Zufallsauswahl stammen, deren Mittelwert und Standardabweichung bekannt sind. Es kann eine Obergrenze formuliert und der Teil an Werten identifiziert werden, der mehr als eine bestimmte Anzahl von Standardabweichungen vom Mittelwert entfernt liegt. Die einfachste Formulierung kommt ohne Annahme über die Form der Verteilung aus:

Bienaymé-Tschebyscheff-Ungleichung:

$$P(|x-a| > t\sigma) \leq 1/t^2$$

$E(x)=a$ und $E(x-a)^2=\sigma^2$ existieren und $t>1$ ist.

Für jede Wahrscheinlichkeitsverteilung ist die Wahrscheinlichkeit, daß eine Variable mehr als $t\sigma$ von ihrem Mittelwert abweicht, höchstens $1/t^2$.

Die Technik besteht nun darin, einen beobachteten Wert mit Hilfe des bekannten μ und σ zu standardisieren, den Ausdruck zu quadrieren und den reziproken Wert zu bilden. Die Ergebnisse gelten als sehr grobe Schätzungen des Ungewöhnlichkeitsmaßes eines Wertes.

Es gibt Verallgemeinerungen dieser einfachen Ungleichung. Sie erfordern mehr Informationen über die Verteilungseigenschaften, z.B. höhere Momente der Verteilung. Das sind Ungleichungen "vom Tschebyscheff - Typ". Sie sind wesentlich wirkungsvoller, allerdings auch sehr viel schwieriger zu berechnen. Während man die Tatsache, daß μ und σ bekannt sein müssen, als theoretischen Nachteil werten kann, gilt die relativ unscharfe Begrenzung als erheblicher praktischer Nachteil. In Situationen, in denen Informationen über Mittelwert und Standardabweichung bekannt sind, nicht jedoch die Variabilitätsstruktur oder die Spannweite, stellen die Verfahren auf der Grundlage von Tschebyscheff-Ungleichungen eine grobe Methode für die Messung der Ungewöhnlichkeit der Daten dar. Parametrische Tests, aber auch das vorher beschriebene Verfahren auf der Grundlage einer bekannten Spannweite, sind den hier genannten Ungleichungen überlegen.

33 Deterministische Kontrollen

Diese Kontrollen sind dadurch gekennzeichnet, daß den Einzelwerten oder Wertekombinationen, die sie durchlaufen, keine differenzierten Maße für ihre Ungewöhnlichkeit (Wahrscheinlichkeitsmaße) zugeordnet werden, sondern - gewissermaßen - die Grenzwerte 0 bzw. 1; das ist gleichbedeutend mit ihrer Rückweisung oder Annahme. Das setzt voraus, daß bei der Gestaltung des Kontrollrahmens alle Ausschlüsse und Zulassungen einzeln zu formulieren sind.

Gelegentlich wird für diese Kontrollen auch der Begriff "Wenn-Dann"-Kontrollen verwendet. Damit sind nicht nur bivariate Fälle gemeint, wie das folgende Beispiel eines trivariaten Falles aus dem Merkmalskatalog einer Volkszählung zeigt: Wenn das Alter "unter 16 Jahre", das Geschlecht "weiblich", der Familienstand "verheiratet" ist, dann ist (zumindest für deutsche Staatsangehörige) die Wertekombination zu beanstanden.

Die Entscheidung über den Annahme- oder Rückweisungsbereich kann auf falschen Voraussetzungen beruhen. Wenn im genannten Fall die Staatsangehörigkeit in die Kontrolle nicht einbezogen wird, dann erfolgt eine Reihe von Rückweisungen zu Unrecht, da bei Ausländerinnen ein Heiratsalter unter 16 Jahren vorkommen kann. Dies kann als Fehler bei der Konzeption des Kontrollrahmens angesehen werden. Generell erfordert die Feststellung ungültiger Werte die Kenntnis historischer Häufigkeitsverteilungen vor allem an den Rändern. Diese allein reicht aber nicht aus, wenn im Zeitablauf bis zur aktuellen Erhebung Strukturveränderungen eingetreten sind. Ein Zuviel an Rückweisungen wegen zu vorsichtiger Fixierung der Kontrollgrenzen ist bei Integration von Korrekturen in den automatisch durchgeführten Prozeß der Fehlerbereinigung im Hinblick auf den Zeitaufwand nicht problematisch; bedenklicher ist, wenn die Beseitigung beträchtlicher Fehleranteile infolge eines nicht ausgereiften Korrekturkonzepts zu Verfälschungen der Ergebnisse oder zu umfangreicher manueller Nachbearbeitung führt.

Bevor die besondere Ausgestaltung deterministischer Kontrollen bei Volkszählungen untersucht wird, werden diese allgemein entwickelt.

331 Funktionale Kontrolle

Mit dieser Kontrolle werden arithmetische Begriffe, z.B. die Summe von Variablen, darauf geprüft, ob ihre Werte zwischen vorformulierten Kontrollgrenzen liegen. Das allgemeine Kriterium für die Annahme eines Datensatzes, der dieser Kontrolle unterworfen wird, lautet:

$$G_u < f(x_1, x_2, \dots, x_n) \leq G_o$$

G_u : Untergrenze

G_o : Obergrenze

$x_1, x_2, \dots, x_n$: kontrollierte Werte

Ein wichtiger Spezialfall ist der univariate Fall, d.h. die Funktion beinhaltet nur eine Variable. Weitere Spezialfälle sind die Additions- bzw. Nullkontrolle und die Quotientenkontrolle.

3311 Additions- bzw. Nullkontrolle

Für diese Kontrolle gilt:

$$\begin{aligned} x_1 + x_2 + x_3 \dots + x_n &= \sum x_i \quad (\text{Additionskontrolle}) \\ \text{bzw. } x_1 + x_2 + x_3 \dots - \sum x_i &= 0 \quad (\text{Nullkontrolle}) \end{aligned}$$

Diese Gleichungen werden auch Identitätsgleichungen genannt. Sie wurden früher in erster Linie als technische Kontrollen eingesetzt, werden heute jedoch auch zur Kontrolle von Datensätzen in der Statistik verwendet.

Dazu ein Beispiel aus der Landwirtschaftszählung 1971 in der Bundesrepublik:

Datenfelder 62 - 72 : Positionen landwirtschaftlich genutzter Fläche

Datenfeld 73 : landwirtschaftlich genutzte Fläche insgesamt
 $\sum W(i) = S \quad (S > 0) \quad W: \text{Wert}$
 $i = 62-72$

Es waren drei Kontrollfragen vorgesehen, deren Verletzung zu Rückweisungen führte:

1. $W(73) = 0$
2. $|W(73) - S| = D,$ D: Differenz
 $[D < 0,1 \cdot W(73)] \vee [D < 1ha]$
3. $W(73) - S = D,$
 $[D \geq 0,1 \cdot W(73)] \vee [D \geq 1ha]$

Fehler sind gegeben, wenn bei positiver Summe der Einzelwerte der Gesamtwert in Feld 73 null ist. Im Falle von Differenzen sind, zum Zwecke unterschiedlicher Behandlung bei der Korrektur, zwei Bedingungen formuliert. Die erste betrifft "Kleinstabweichungen" mit einer Differenz von weniger als 1/10 der Summe der Einzelpositionen oder weniger als 1 ha und die zweite mindestens 1/10 der Summe der Einzelpositionen oder mindestens 1 ha.

3312 Quotienten-Kontrolle

Mit dieser Kontrolle werden statistische Beziehungszahlen darauf geprüft, ob ihre Werte in einem bestimmten, aufgrund z.B. historischer Kenntnisse vorformulierten, Annahmebereich liegen.

Für diese Kontrollen gilt allgemein:

$$G_u < R \leq G_o$$

Ein wichtiger Anwendungsbereich für diese Kontrolle ist bei der Außenhandelsstatistik gegeben, wo für jede Warenart die Relation Wert zu Menge getestet werden kann.

Dieses Verfahren kann dadurch verfeinert werden, daß der Zähler und/oder der Nenner zusätzlich zum Quotienten gegen bestimmte Grenzen geprüft werden. Außerdem können historische Vergleiche durchgeführt werden dadurch, daß Daten aus dem aktuellen Berichtszeitraum auf ältere bezogen werden. Wenn Mengen und Werte nach Warengruppen aggregiert und durchschnittliche Quotienten aus diesen Aggregaten berechnet werden, dann können solche Quotienten möglicherweise engeren Grenzen unterworfen werden (bedingte Kontrollen).

3313 Anwendung funktionaler Kontrollen auf Volkszählungen

Um die Eignung funktionaler Kontrollen für Volkszählungen zu untersuchen, muß näher auf die Art der Merkmale bei Volkszählungen und ihre Skalierungseigenschaften eingegangen werden. Außerdem ist zu prüfen, ob die Informationsredundanz besondere Möglichkeiten der Fehleridentifikation bietet.

33131 Art der Merkmale bei Volkszählungen

Die Gestaltung des Kontrollrahmens hängt ab von den Merkmalsstrukturen, die in einer Volkszählung erfragt werden. "Klassische" Merkmale einer Volkszählung sind "Geschlecht", "Familienstand", "Religion", "Staatsangehörigkeit", "Alter", "Beruf". Mit Ausnahme des Alters handelt es sich hierbei nur um qualitative Variable. Auch die Gesamtübersicht über den Merkmalskatalog der Volkszählung 1970 im Anhang¹⁾ zeigt, daß unter den übrigen Merkmalen nur wenige quantitativer Art sind. Fast alle Merkmale werden auf einer Nominalskala "gemessen". Gleiche Ausprägungen werden gleichen Klassen zugeordnet. Dabei gibt es nur gleich bzw. ungleich, aber nicht größer bzw. kleiner. Daher können Mittelwerte und Varianzen nicht berechnet werden. Funktionale Kontrollen, bei denen die Variablen bestimmten arithmetischen Bedingungen unterworfen werden, sind bei Volkszählungen nur im Rahmen der quantitativen Merkmale anwendbar.

33132 Informationsredundanz

Meßwerte aus dem naturwissenschaftlichen Bereich lassen oft keine strukturellen Zusammenhänge erkennen. Sie haben als quantitative Begriffe aber den Vorzug, daß hochentwickelte mathematisch-statistische Verfahren wie etwa die Hauptkomponentenanalyse auf sie angewandt werden können. Dagegen können bei Volkszählungen die in den Informationen enthaltenen Redundanzen zum Zwecke der Fehleridentifikation genutzt werden. Darunter werden "überflüssige Elemente" in den Informationen verstanden.

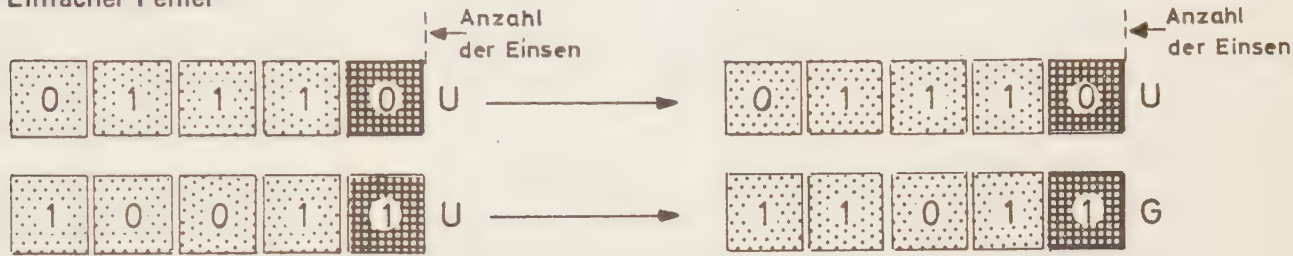
1) Vgl. Fragebogen im Anhang; Gesetz über eine Volks-, Berufs- und Arbeitsstättenzählung (Volkszählungsgesetz 1970) vom 14. April 1969 (BGBl. I S.292).

Wenn aufgrund gesetzlicher Vorschriften die Erwerbstätigkeit von Personen bis zum Alter von 14 oder 15 Jahren ausgeschlossen ist, dann impliziert das Merkmal "Erwerbstätigkeit" diese Altersjahre als Untergrenze. Der Verfasser nennt dies "implizite" Redundanz.

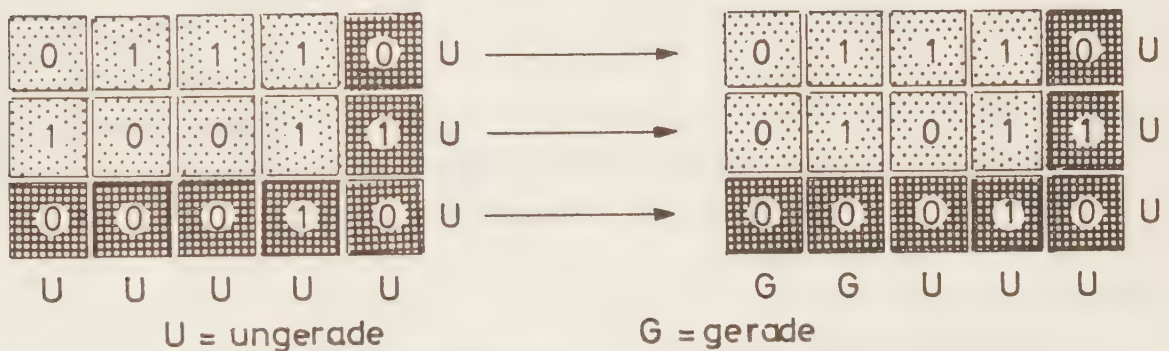
Davon zu unterscheiden sind "explizite" Redundanzen, die bewußt geschaffen werden zur Fehleridentifikation. Beispiele dafür finden sich bei der maschineninternen Kontrolle der Informationsübertragung und -speicherung, etwa bei der Odd-parity-Kontrolle.¹⁾

DIE ODD - PARITY - KONTROLLE

Einfacher Fehler



Doppelter Fehler



Erläuterung des "Einfachen Fehlers"

Links ist der Zustand vor der Datenübertragung abgebildet. Die Dezimalzahlen 7 und 9 sind in einem Halbbyte binär dargestellt ($7 = 0111$ und $9 = 1001$). Diesen Bitkombinationen ist als 5. Bit ein Prüfbit zugeordnet. Es enthält den Wert 1 oder 0 so, daß die Anzahl der Einsen in der gesamten Informationseinheit

1) Brennan, J. F., Computer "reparieren" sich selbst - Entwicklung von Verfahren zur automatischen Fehlersuche und -korrektur, in: IBM Nachrichten, 20.Jg., H. 199, Verlag IBM Deutschland, Sindelfingen 1970, S.7-13.

immer ungerade ist. Rechts ist der Zustand nach der Datenübertragung abgebildet. Bei der Dezimalzahl 7 hat sich nichts geändert, bei der Übertragung der 9 ist in einem Bit fälschlich eine Umwandlung von 0 in 1 erfolgt. Die nunmehr entstandene gerade Zahl von Einsen führt zur Fehleridentifikation.

Erläuterung des "Doppelten Fehlers"

Da sich bei zwei Übertragungsfehlern in einer Informationseinheit die Anzahl der Einsen nicht von ungerade in gerade verändern würde, kann zusätzlich zur horizontalen noch eine vertikale Bitprüfung vorgenommen werden. Die Anzahl der Einsen einschließlich des Prüfbits müßte also in der Zeile und in der Spalte ungerade sein. Die Abbildung zeigt zwei Übertragungsfehler bei der Dezimalzahl 9 in den ersten beiden Bits.

Für die beiden häufigsten Fehlertypen - eine Ziffer des Codes ist falsch bzw. zwei aufeinander folgende Ziffern sind vertauscht - hat Bose ein Rechenverfahren entwickelt, mit dessen Hilfe aus sechsstelligen Kennziffern einstellige Kontrollziffern abgeleitet werden können.¹⁾ Die künstliche Schaffung von Redundanzen ist bisher sehr erfolgreich z.B. für Versicherungs- und Schecknummern angewandt worden.

Die von der amtlichen Statistik in der Bundesrepublik benutzten Schlüsselverzeichnisse enthalten keine Kennziffern mit Kontrollbegriffen. Bei einer Volkszählung wäre der Aufwand, für Millionen von Ziffern zusätzliche Kontrollziffern zu schreiben, untragbar. Außerdem ist die Gefahr der Fehlzuordnung eines Schlüssels zu einer Merkmalsausprägung wesentlich höher zu bewerten als die Verfälschung eines Schlüssels. Explizite Redundanzen in Form bewußt in den Fragenkatalog aufgenommener Kontrollfragen sind unter Kostengesichtspunkten sicherlich kaum durchzusetzen. Wenn implizite

1) Bose, R. C., Error Detecting and Error Correcting Indexing Systems for Large Serial Numbers, in: Review of the International Statistical Institute, Vol. XXXIV, Book 3, The Hague 1966, S.334. Eine Kurzfassung dieses Artikels befindet sich in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol.XLI, Book 1, Beograd 1966, S.472.

Redundanzen zur Fehleridentifikation, besonders aber zur Abschätzung von Fehlerwahrscheinlichkeiten in Merkmalskombinationen (Fehlerallokation), in einzelnen Merkmalsbereichen nicht gegeben sind, dann müssen größere Unsicherheiten in Kauf genommen werden.

332 Automatische Signierkontrollen

Das sind Kontrollen des Signierprozesses, in dem verbale Informationen in Codes transformiert werden. Nach Einführung der optischen Beleglesung kann die Umsetzung bereits vorhandener numerischer Begriffe in maschinell lesbare stilisierte Ziffern dazugerechnet werden. Im Rahmen der automatischen Fehlerbereinigung ist diese Kontrolle als Totalkontrolle angelegt; sie findet sich als Stichprobenkontrolle aber bereits im Vorfeld der maschinellen Datenverarbeitung.¹⁾

Die Signierkontrolle ist eine Einzelfeld-Kontrolle und im Normalfall univariat. Seltener treten in einem Datenfeld Signierbegriffe auf, die aus mehreren Merkmalen zusammengesetzt sind (multivariater Fall). Ein synonyme Begriff für Signierkontrolle ist "Bereichskontrolle" (Range Check). Nach Transformation der Informationen zum Geburtsmonat (Januar bis Dezember) in die Ziffern 01 bis 12 wird das zutreffende Datenfeld im Datensatz abgefragt und der Wertebereich zwischen 01 und 12 kontrolliert. Ein anderes Beispiel betrifft die Ausprägungen des Merkmals "Familienstand". Es sei folgender Code vereinbart:

ledig	1
verheiratet	2
verwitwet	3
geschieden	4

Hier umfaßt der Wertebereich die Ziffern 1 bis 4. Da das Merkmal "Alter" bei etwa 110 Jahren eine "natürliche Begrenzung" aufweist, kann das Datenfeld, in dem dieses Merkmal definiert ist, auf Werte von 001 bis 110 geprüft werden. In einer speziell durchge-

1) Vgl. Abschnitt 34 "Statistische Qualitätskontrolle" S.55 ff.

führten Untersuchung weist der Verfasser nach, welche Fehlerursachen eine Verfälschung des Alters haben kann.¹⁾ Ein wesentlich umfangreicheres Schlüsselsystem ist das der Berufsbenennungen. Es umfaßte bei der Volkszählung 1970 mehrere tausend Codes. Bei systematisch aufgebauten Schlüsseln (das wird auch "mnemotechnisches Prinzip" genannt) gibt es unverbundene Bereiche im Schlüsselverzeichnis.

So ist z.B. innerhalb der Berufsordnung "Ärzte" (841) die Benennung "Urologe" (841-061) der letzte vorkommende Signierbegriff. Der nächste gültige Schlüssel liegt innerhalb der Berufsordnung "Zahnärzte" (842) mit der Benennung "Kieferorthopäde" (842-001). Von den rd. 1 Mill. Schlüsseln, die bei einem sechsstelligen numerischen Begriff kontinuierlich vergeben werden könnten, waren in diesem System nur 19 203 (1,92 %) besetzt.

Die gültigen oder alternativ die ungültigen Kennziffern werden als Verzeichnis ("Tabelle") gespeichert. Umfangreiche Tabellen befinden sich meist auf "Leitbändern". Der Kontrollvorgang besteht darin, den zu prüfenden Schlüssel mit dem Schlüssel dieser Tabelle zu vergleichen. Es handelt sich dabei um eine Prüfung auf formale Fehler. Die Abfrage kann programmtechnisch kontinuierlich oder in Sprüngen (z.B. Intervall-Halbierungsverfahren oder Dreier-sprünge usw.) erfolgen. Die Sprungverfahren erfordern bei umfangreichen Tabellen weniger Maschinenzeit.

Zum Schluß ein Beispiel für einen Signierbegriff, der aus einer Merkmalskombination besteht. Es handelt sich um einen Ausschnitt aus dem bei der Volkszählung 1961 benutzten Schlüsselverzeichnis für die Kombination der Merkmale "Stellung im Beruf" und "Soziale Stellung":

Beamter

des einfachen und mittleren Dienstes	60
des gehobenen Dienstes	61
des höheren Dienstes	62
Angabe der Laufbahngruppe fehlt	69

1) Vgl. S.147/148.

3321 Kombinationskontrollen

Kombinationskontrollen können als Signierkontrollen mehrerer Datenfelder "gleichzeitig" angesehen werden. Diese in einem für sinnvoll gehaltenen Zusammenhang stehenden Datenfelder eines Datensatzes (z.B. Einzelperson) oder auch mehrerer Datensätze (z.B. Haushalt) werden nicht bereits in einem Schlüsselverzeichnis zusammengesetzt, sondern erst im Kernspeicher der Datenverarbeitungsanlage, wobei nacheinander ablaufende Verarbeitungsschritte wegen der sehr großen Geschwindigkeiten praktisch als "gleichzeitig" betrachtet werden können. In diesem Sinne ist die Kontrolle der o.g. Merkmalskombination eine Kombinationskontrolle. In der Literatur werden auch quantitative Merkmale in Volkszählungen wie "Zahl der Kinder", "Alter" und "Ausbildungsdauer" als Codes aufgefaßt und - im Falle ihrer Verknüpfung - "Kombinationskontrollen" unterworfen.¹⁾ In Anbetracht des Datencharakters schlägt der Verfasser vor, in diesem Fall von "funktionalen Kontrollen" zu sprechen. Das trifft z.B. auf die Überprüfung des Generationsabstandes im Kontrollrahmen der Volkszählung 1970 zu. Dabei wurde das Alter des Haushaltsvorstandes bzw. seines Ehegatten mit dem des ältesten im Haushalt lebenden ledigen Kindes verglichen, der bezüglich des männlichen Haushaltsvorstandes auf mindestens 18 Jahre, bezüglich des weiblichen Haushaltsvorstandes auf mindestens 16 Jahre determiniert war.

Bei der Konzeption des Kontrollrahmens ist zu bedenken, daß die Bewertung einer fehlenden Information (Blank) als teilweiser Informationsausfall oder als zu Recht weggelassen nur aus der Verknüpfung von Datenfeldern heraus zu beurteilen ist. So stellt etwa ein fehlender Berufsschlüssel eine korrekte Information für einen Nichterwerbstätigen unter 15 Jahre dar.

3322 Binärkettentechnik (binary string)

Bei umfangreichen Schlüsselsystemen, bei denen die Codes aufgrund des systematischen Aufbaus noch dazu unverbunden sind, ist die

1) Vgl. Nordbotten, S., Automatic Editing of Individual Statistical Observations, a.a.O., S.15-17.

Binärkettentechnik¹⁾ - vor allem angesichts der beschränkten Speicherkapazität der Datenverarbeitungsanlagen der zweiten Computergeneration - besonders vorteilhaft gewesen. Diese Ketten bestehen aus jeweils so vielen Positionen wie Schlüssel möglich sind. Jeder gültige Schlüssel kann mit binär eins, jeder ungültige mit binär null gekennzeichnet werden. Folgender Ausschnitt aus einem systematisch aufgebauten Schlüsselverzeichnis soll das Vorgehen verdeutlichen:

Signierschlüssel

-	-	-	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	-	-	-

Gültige Kennziffern seien die nicht durch Kreuz markierten zwischen 13 - 36; die markierten seien ungültig. Die Ziffern 13 und 36 sowie eine Binärkette mit 24 Positionen werden gespeichert. Es wird kontrolliert, ob ein Schlüssel in dem Intervall liegt und ob - bei aufsteigender Reihenfolge - in seiner Position binär eins für einen gültigen Code gespeichert ist. Je nach Aufbau des Schlüssels ist diese Methode für einzelne Merkmale wie für Merkmalskombinationen anzuwenden.

3323 Zukunft automatischer Signierkontrollen

Als Folge der Substitution des Ablochprozesses durch die optische Beleglesung wird die automatische Signierkontrolle für Merkmale mit vorgegebenen Ausprägungen überflüssig. Bei umfangreichen Schlüsselssystemen (Gemeindekennziffern, Wirtschaftszweige, Berufe, Staatsangehörigkeiten, Religionsgemeinschaften) wird sie jedoch weiterhin benötigt.

Eine andere Entwicklung, die ebenfalls die Einschränkung von Signierkontrollen zur Folge hat, schreitet voran: die maschinelle Signierung. Das setzt den Aufbau und die Speicherung von Verzeichnissen voraus, die Wortadressen in Verbindung mit numerischen Codes enthalten. Die von den Auskunftspersonen in die Zählpapiere eingetragenen Texte werden in der Datenverarbeitungsanlage identifiziert, und bei Übereinstimmung wird der

1) Vgl. Kivivuori, R., A Method for Checking Numerical Codes Using the 1401, in: Nordisk Tidskrift for informasjonsbehandling, Bd 1, H. 1, Lund 1961, S.48-53.

numerische Code übernommen. Gegenwärtig werden Maschinenprogramme entwickelt, mit denen es möglich sein soll, auch nicht-standardisierte Texte mit den Adressen des Leitbandes zusammenzuführen. Die Übereinstimmung wird nach zwei Prinzipien festgestellt: Ganzwort-Vergleich oder Vergleich von Adreß-Komponenten mit Vergabe von Punkten.¹⁾ Die höchste Punktzahl entscheidet über die Zuordnung. In USA sind solche Systembänder bereits für die Codierung von Regionalangaben und diejenige wirtschaftlicher Merkmale von Industriebetrieben vorhanden.²⁾

333 Mengenansatz von Fellegi/Holt³⁾

Hierbei handelt es sich um Fehleridentifikation in dem Sinn, daß ein Minimum an Datenfeldern eines Datensatzes determiniert wird, das bei "sinnvoller" Änderung der in diesen Feldern befindlichen Werte ausreicht, damit alle Kontrollen des Prüfrahmens unbeanstandet durchlaufen werden.

Der Lösungsansatz geht davon aus, daß jede Kontrolle eine logische Verknüpfung von Wertemengen bestimmter Datenfelder eines Datensatzes ist. Er beschränkt sich nicht auf quantitative Daten und ist daher im Prinzip auf Volkszählungen anwendbar. Wertemengen sind die Ausprägungen jedes Merkmals, die bei einer Verknüpfung von Merkmalen einzeln als für die betreffende Kontrolle gültig angegeben sein müssen. Kontrollen, die zu Beginn des Fehlerbereinigungsprozesses spezifiziert werden, heißen hier "explizite" Kontrollen.

Als "implizite" Kontrollen werden diejenigen bezeichnet, die in einem Datenfeld Wertemengen aufweisen, die im selben Datenfeld einer expliziten Kontrolle enthalten sind. Wenn ein zu kontrollie-

-
- 1) Wichtige Grundsätze bei der Messung der "größten Ähnlichkeit" in der Textverarbeitung hat Thorelli dargestellt; Thorelli, L. E., Automatic Correction of Errors in Text, in: Nordisk Tidskrift for informasjonsbehandling, Bd 2, H. 1, Lund 1962, S.45-52.
 - 2) Minton, G., Inspection and Correction Error in Data Processing, in: Journal of the American Statistical Association, Vol. 64, No. 328, Menasha, Wisc. 1969, S.1272.
 - 3) Fellegi, I. P., Holt, D., A Systematic Approach to Automatic Editing and Imputation, in: Bulletin of the International Statistical Institute, Proceedings of the 39th Session, Vol.44, Book 1, Wien 1973, S.364-376.

render Datensatz also eine implizite Kontrolle verletzt, verletzt er unbedingt auch eine explizite Kontrolle. Wenn die Wertemenge in dem Datenfeld einer impliziten Kontrolle mit der Wertemenge in demselben Datenfeld einer expliziten Kontrolle vollständig übereinstimmt, dann kann sie in der impliziten Kontrolle herausfallen, weil sie redundant ist, d.h. kein zusätzliches Kontrollkriterium darstellt. Diese implizite Kontrolle heißt dann "wesentlich neue implizite" Kontrolle.

Die Menge von Kontrollen heißt "vollständig", wenn die expliziten Kontrollen um alle wesentlich neuen impliziten Kontrollen, die aus ihr abgeleitet werden können, erweitert werden. Daraus leitet sich umgekehrt das Konstruktionsprinzip für die vollständige Menge von Kontrollen ab. Damit sind alle Bedingungen für Kontrollverletzungen spezifiziert.

Mit Hilfe von Mengenoperationen behaupten und beweisen die Autoren ein Lemma, das ein allgemeines Verfahren zur Bestimmung impliziter Kontrollen zum Inhalt hat. Die Operation ist vorteilhaft zur Aufdeckung interner Inkonsistenzen in den expliziten Kontrollen.

Ein Theorem der Autoren besagt: Wenn a_i ($i = 1, 2, \dots, K-1$) mögliche Werte für $K-1$ Datenfelder eines Datensatzes sind und diese Werte alle Kontrollen mit Verknüpfungen der Datenfelder $1, \dots, K-1$ erfüllen, dann gibt es einen Wert a_K so, daß die Werte a_i ($i = 1, 2, \dots, K$) alle Kontrollen mit Verknüpfungen der Datenfelder $1, \dots, K$ erfüllen. Das Theorem ist nur gültig, wenn die Menge der Kontrollen vollständig ist.

Durch wiederholte Anwendung dieses Theorems können Werte für die Felder $a_{K+1}, \dots, a_N$ gefunden werden (N ist der komplette Datensatz) so, daß alle Kontrollen mit Verknüpfungen der Datenfelder $1, \dots, N$ erfüllt werden.

Wenn nun Mengen von Datenfeldern ausgewählt werden so, daß wenigstens ein Feld in allen verletzten Kontrollen vorkommt, dann existiert eine Wertemenge, die - zusammen mit den Wertemengen anderer Datenfelder - allen Kontrollen genügt. Die Wahl dieser Datenfelder ergibt jedoch nicht zwangsläufig die kleinstmögliche Anzahl. Die Autoren schlagen vor, mit der kleinstmöglichen

Zahl von Datenfeldern dann zu arbeiten, wenn die Wahrscheinlichkeit für Fehler in einzelnen Datenfeldern gering ist. Durch "sinnvolle" Änderung der Inhalte dieser Felder kann erreicht werden, daß alle Werte eines Datensatzes die Gesamtheit der Kontrollen unbeanstandet passieren.¹⁾

Interessant zu bemerken ist, daß Naus - ohne die Arbeit von Fellegi/Holt zu erwähnen - eine fast übereinstimmende Lösung, jedoch unter Verwendung anderer Begriffe - vorschlägt. Der Verfasser nimmt im nächsten Kapitel dazu Stellung.²⁾

Die Autoren versprechen sich eine wesentliche Vereinfachung der Programmierung durch ihr Verfahren. Komplexe, ineinander verschachtelte Kontrollen, könnten durch einfache Kontrollregeln ersetzt werden. Nach Meinung des Verfassers dürfte jedoch bei Volkszählungen die Zahl impliziter Kontrollen so groß sein, daß selbst bei der erheblichen Speicherkapazität und hohen Verarbeitungsgeschwindigkeit moderner Computer der Suchaufwand zur Bestimmung der kleinsten Felderzahl (und ihrer durch keine Kontrolle mehr zu beanstandenden Inhalte) eine praktische Anwendung des Verfahrens nicht möglich sein wird. Dann enthält die Arbeit aber immer noch eine ausgezeichnete Anleitung zur Analyse deterministischer Kontrollen.

334 Konstruktionsprinzipien des Kontrollrahmens

An einen Kontrollrahmen sind drei Forderungen zu stellen: er sollte vollständig, konsistent und frei von Redundanzen sein.

Fragebögen, die bei Volkszählungen Verwendung finden, enthalten oft 50 und mehr Merkmale mit zum Teil sehr vielen Ausprägungen (z.B. beim Merkmal "Beruf"). Ein Kontrollrahmen, der alle theoretisch möglichen Wertekombinationen umfaßt, ist praktisch nicht zu verwirklichen. Es muß daher eine Beschränkung auf "sinnvolle" Kombinationen erfolgen. Kriterium sollte in erster Linie die Struktur

1) Vgl. S.98.

2) Vgl. S.78.

der Ergebnistabellen sein. Dem Einwand, daß das Standard-Veröffentlichungsprogramm nicht stark gegliedert zu sein braucht, wenn jederzeit Spezialtabellen auf Abruf erstellt werden können, ist entgegenzuhalten, daß solche Tabellen kaum aus den Einzelinformationen entwickelt werden, sondern aus bereits verdichtetem Datenmaterial. Daher muß vor der Konzeption des Kontrollrahmens doch eine Abstimmung erfolgen über die Merkmalskombinationen, die zu Auskünften zur Verfügung stehen sollen und daher in die Fehlerbereinigung einzubeziehen sind. Jede nicht kontrollierte Merkmalskombination kann zu inkonsistenten Informationen führen. Die Lösung von Fellegi/Holt, die eine Vorherbestimmung fachlich sinnvoller Kombinationen nicht erfordert, ist auf ihre Praktikabilität hin zu untersuchen. Sie erweckt den Anschein einer zeit- und kostenintensiven Methode.

Konsistenz als Forderung an die Beschaffenheit der Daten setzt ihre Verwirklichung im Kontrollrahmen voraus. Z.B. darf ein für den univariaten Fall formulierter Annahmebereich bezüglich dieses Merkmals im höhervariaten Fall nicht erweitert werden. Je größer die Datenmengen sind, je mehr Bedingungen infolge verbesserten Wissensstandes den Daten im Laufe der Zeit auferlegt werden, umso wichtiger ist eine ständige Überwachung des Kontrollrahmens bzw. seine gelegentliche Umstrukturierung.

Die Forderung nach Vermeidung von Redundanzen (hier als "explizite" Redundanzen verstanden) ist zwiespältig. Sie muß erhoben werden, um Programmieraufwand und Maschinenlaufzeit zu vermindern. Andererseits zeigt gerade der Abschnitt über die Fehlerallokation, wie wichtig Redundanzen für die Verbesserung der Datenqualität sind.

34 Statistische Qualitätskontrolle

Maschinelle und manuelle Bearbeitung der Belege einer Volkszählung tragen Merkmale industrieller Produktion, so daß die mathematisch-statistischen Verfahren zur laufenden Qualitätsüberwachung der Arbeitsgänge (Prozeßkontrolle) bzw. zur Qualitätsbeurteilung der Arbeitsmassen (Annahmestichproben) hier Anwendung finden können.¹⁾ Den Besonderheiten statistischer Arbeiten wird durch die "hybriden" Kontrollen (das ist eine Kombination beider Verfahren) zusätzlich Rechnung getragen. Ziel Statistischer Qualitätskontrollen ist es, Qualitätsstandards und Produktions- sowie Kontrollkosten aufeinander abzustimmen.

Im Gegensatz zur automatischen Fehlerbereinigung werden die Daten bei der Statistischen Qualitätskontrolle nur stichprobenhaft geprüft und korrigiert. Wenn zurückgewiesene Arbeitsmassen vollständig überarbeitet werden, erhöht sich der Anteil des insgesamt bereinigten Materials. Fehler der Datenerfassung, hier interpretiert als unvollständige Identifikation der statistischen Gesamtheit und Signierfehler als Fehlzuordnung der statistischen Einheiten in das Klassifikationssystem, können substantiell nur durch Nacherhebung bzw. korrekte Codierung verbessert werden. Automatische Fehlerbeseitigung bei Bestandsverlusten durch Doppelung von Datensätzen kann nur im äußersten Notfall als Ausweg gelten, automatische Korrekturen von Signierbegriffen sind nur imstande, formale "Richtigkeit" herzustellen. Wenn durch maschinellen Vergleich des Codes mit dem gespeicherten Schlüsselverzeichnis keine Beanstandung festgestellt wird, weil das Schlüsselssystem zu kompakt ist und jede Verfälschung einen gültigen Begriff erzeugt,²⁾ dann sind Qualitätsverbesserungen nur durch die Statistische Qualitätskontrolle zu erreichen. Für die Lochung der Belege und die Beleglesung gilt dasselbe. Alle genannten Arbeitsgänge befinden sich im Vorfeld der automatischen Datenverarbeitung.

1) Vgl. Uhlmann, W., Statistische Qualitätskontrolle, Teubner, Stuttgart 1966; Sauter, U., Die moderne statistische Qualitätskontrolle unter besonderer Berücksichtigung ihrer Bedeutung für die Wirtschaftlichkeit der Unternehmung, Diss., Mannheim 1959.

2) Vgl. Schuff, H. K., Fehler in informationsverarbeitenden Systemen und ihre Bedeutung, Teil II, in: Elektronische Datenverarbeitung, H. 2, Braunschweig 1964, S.67.

341 Prozeßkontrolle

Das wichtigste Hilfsmittel der Prozeßkontrolle ist die Kontrollkarte, in die entweder die Meßwerte der geprüften Merkmale von laufend durch Zufallsauswahl entnommenen Einheiten oder Fehleranteile eingetragen werden. In der Statistischen Qualitätskontrolle wird grundsätzlich zwischen der Variablenkontrolle und der Attributenkontrolle (Gut-Schlecht-Prüfung) unterschieden. Ausgehend von bestimmten Annahmen über die Verteilung der Grundgesamtheit (bei quantitativen Variablen Normalverteilung, bei (0,1)-verteilten Variablen Binomial- oder Poissonverteilung) werden die die Verteilung charakterisierenden Parameter (Mittelwert, Varianz oder Spannweite) als "Sollwert" bzw. Kontrollgrenzen auf den Karten markiert. Meist wird zusätzlich zu den Kontrollgrenzen ein Toleranzbereich mit größerem Intervall auf Grund technischer Voraussetzungen gekennzeichnet. Wenn die Testgrößen in den Kontrollbereich fallen, dann werden die Schwankungen um den Sollwert als zufällig angesehen; der Prozeß gilt als "unter statistischer Kontrolle" stehend. Werte außerhalb dieses Bereiches (das entspricht der Ablehnung der Nullhypothese in der Testtheorie) bedeuten signifikante Abweichungen, die ungewöhnliche Einflüsse signalisieren können. Dadurch, daß die Ursache für die Störung erkannt und der Fehler beseitigt wird, kann der Prozeß wieder "beherrscht" werden. Das kann problemlos sein, wenn es sich um eine Maschine handelt, die neu justiert oder wegen Materialabnutzung ersetzt werden muß. Beispielhaft dafür sind im US Census die Mikrofilm-Kameras bei der optischen Erfassung der Informationen von den Belegen, bei denen signifikante Abweichungen von den Sollabständen zwischen den Einzelaufnahmen vorkommen können.¹⁾ Da jedoch die manuellen Arbeiten bei weitem überwiegen, liegt das Schwergewicht auf der Qualitätskontrolle der Arbeitsleistung des Mitarbeiterstabes. Fehleranteile außerhalb der Kontrollbereiche, die in Abhängigkeit vom jeweiligen Schwierigkeitsgrad unterschiedlich spezifiziert waren, führen hier zu erneuten Unterweisungen oder zur Ablösung, wenn die Qualitätsanforderungen eine Zeitlang nicht erfüllt werden können.

1) Vgl. United States Censuses of Population and Housing 1960, Quality Control of Preparatory Operations, Microfilming, and Coding, Hrsg. US Bureau of the Census, Washington, D.C. 1965.

342 Annahmestichproben

Die Beurteilung der Qualität von Arbeitsmassen nach den Prinzipien Statistischer Qualitätskontrolle setzt die Konzeption von Stichprobenplänen voraus. Es wird unterschieden zwischen einfachen und mehrfachen bzw. sequentiellen Stichprobenplänen. Während die Entscheidung auf der Grundlage eines einfachen Planes nach einmaligem Ziehen einer Zufallsstichprobe getroffen wird, können bei mehrfachen Plänen mehrere Stichproben notwendig werden, um zur Entscheidung über Annahme oder Rückweisung einer Arbeitsmasse zu gelangen. Das Prinzip der Kontrolle läßt sich am einfachen Stichprobenplan erläutern; der mehrfache Plan ist schwieriger abzuleiten, weil er mehr Parameter berücksichtigt.

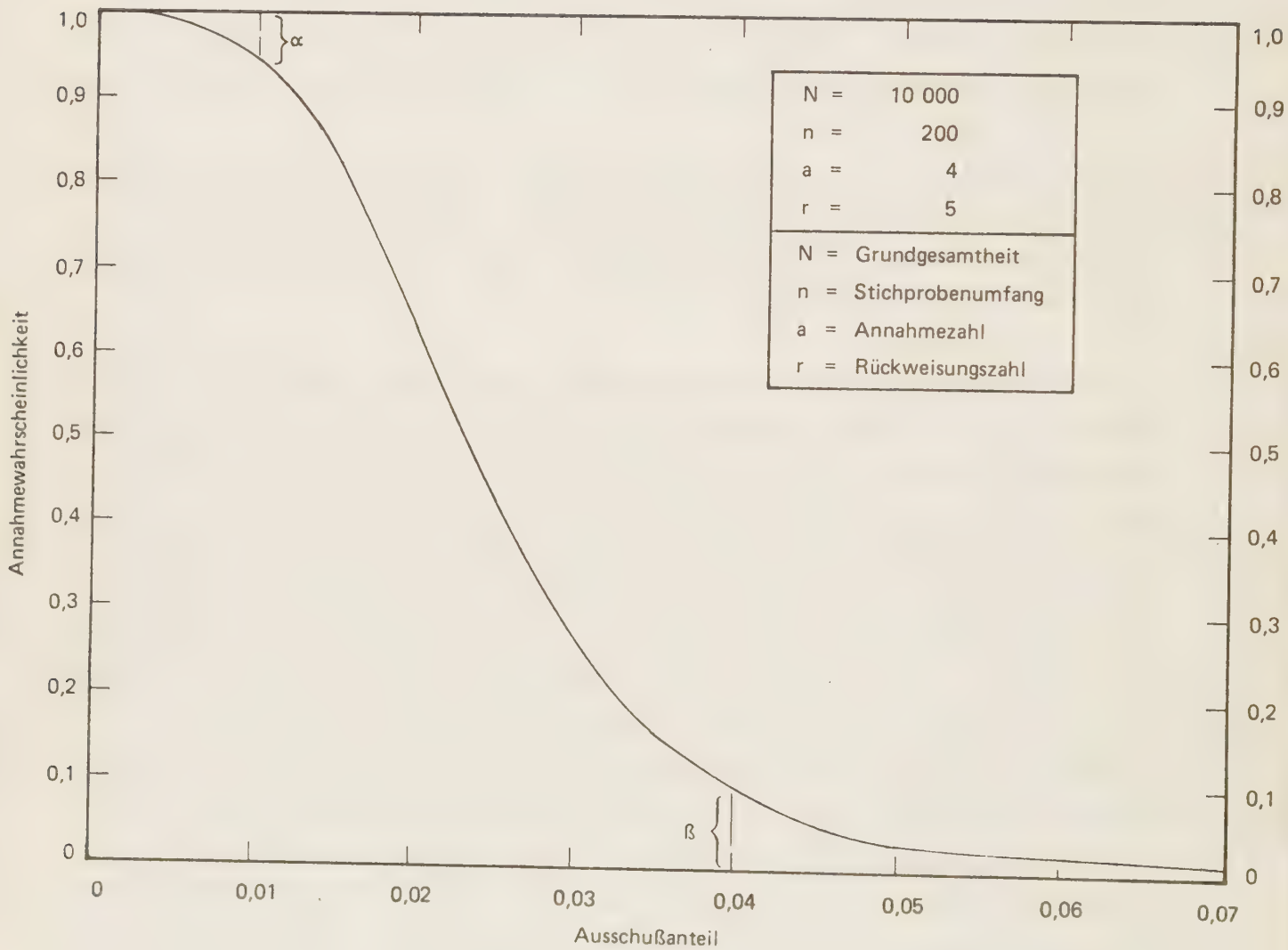
Die Eigenschaften der Stichprobenpläne lassen sich durch zwei Kurven beschreiben: Annahmekennlinien (auch Operationscharakteristiken) und Kurven durchschnittlicher Auslieferungsqualität (mittlerer Durchschlupf).¹⁾

Die Annahmekennlinien sind bestimmt durch den Umfang der Arbeitsmasse (N), den Stichprobenumfang (n) und - bei der Attributenkontrolle - durch die Annahmezahl (a) bzw. Rückweisungszahl (r).²⁾ Das Auffinden von fehlerhaften Einheiten in der Stichprobe bis zum Erreichen der Annahmezahl hat die Annahme der Arbeitsmasse zur Folge. Bei mehr fehlerhaften Einheiten erfolgt die Ablehnung der Arbeitsmasse. Die Funktion der Operationscharakteristik wird abgeleitet aus der Gütefunktion des dem Stichprobenplan zugrunde liegenden Tests. Ein Punkt auf der Annahmekennlinie gibt den funktionalen Zusammenhang zwischen Ausschußanteil und Annahmewahrscheinlichkeit an. Bei der Attributenkontrolle dient die hypergeometrische Verteilung als Berechnungsgrundlage für die Funktion. Wenn n/N sehr klein ist, so daß die Auswahl der Stichprobenelemente die Wahrscheinlichkeiten kaum verändert, kann die Binomial- bzw. Poisson-Verteilung zugrunde gelegt werden. Die Annahmekennlinie muß in ihrer Konstruktion von einem gegebenen Produzentenrisiko α (in der Testtheorie Fehler 1. Art) bei einem

1) In der angelsächsischen Literatur "Average Outgoing Quality" (AOQ).

2) Vgl. Graphik auf der nächsten Seite.

BEISPIEL EINER ANNAHMEKENNLINIE



Quelle: Swoboda, H., Knaurs Buch der modernen Statistik, Droemer Knaur, München-Zürich 1971, S.182.

bestimmten Ausschußanteil und einem Konsumentenrisiko β (entsprechend Fehler 2. Art) bei einem - höheren - Ausschußanteil ausgehen. Die Fähigkeit, zwischen guten und schlechten Arbeitsmassen zu unterscheiden, auch Trennschärfe oder Schärfe der Diskriminierung genannt, kommt in der Gestalt der Operationscharakteristik zum Ausdruck. Die Trennschärfe erhöht sich bei der Vergrößerung des Stichprobenumfangs.

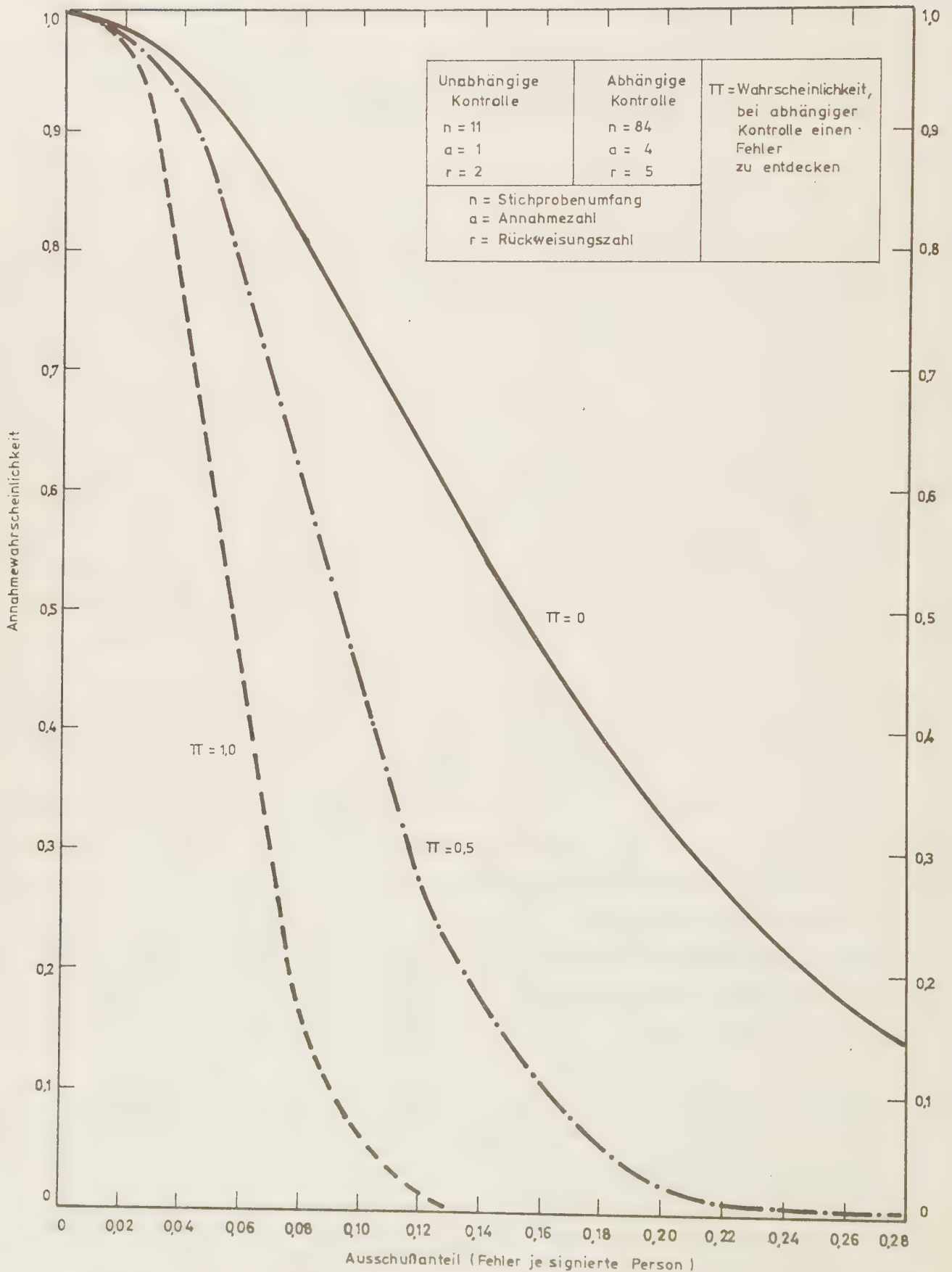
In USA wurden Signierkontrollen nicht erst in der Produktionsphase sondern zu Ausbildungszwecken bereits in einer Übungsphase durchgeführt. Hier wurde die Qualifikation des Signierers für seinen Einsatz in der Produktionsphase getestet. Annahme oder Rückweisung einer Arbeitsmasse hingen vom Ausschußanteil einer "abhängig" geprüften mit einer "unabhängig" geprüften Stichprobe (Stichprobenpaar) ab. Diese Unterscheidung wurde erstmals 1960 getroffen, nachdem deutlich geworden war, daß sich die Prüfer von den Signierentscheidungen der Signierer (ähnlich auch bei der Kontrolle des Lochprozesses) zu stark beeinflussen ließen und so beachtliche Fehleranteile nicht erkannten. Bei der "unabhängigen" Kontrolle wurde zweimal vorsigniert, ehe die endgültige Verschlüsselung stattfand. Über die "Richtigkeit" einer Signierung wurde mit der Mehrheit von 2 : 1 entschieden.

Auf der Graphik der nächsten Seite sind die Annahmekennlinien für das Signieren "allgemeiner" Merkmale in der Übungsphase dargestellt. Unter "allgemeinen" Merkmalen waren mit Ausnahme von "Wirtschaftszweig" und "Beruf" alle übrigen signierbedürftigen zusammengefaßt. Die Kontrolle bezog sich auf insgesamt 10 % aller Haushalte. Unabhängig kontrolliert wurde eine Unterstichprobe mit einem Auswahlatz von 1,25 %. Die Annahmewahrscheinlichkeiten für die Arbeitsmassen wurden alternativ unter der Prämisse berechnet, daß bei der abhängigen Kontrolle a) alle Fehler ($\pi=1,0$), b) die Hälfte der Fehler ($\pi=0,5$), c) kein Fehler ($\pi=0$) entdeckt wurde. Fall b) wurde für realistisch gehalten.

Wenn bei Kontrollen nicht alle Fehler identifiziert werden, ist die Wirksamkeit der Qualitätskontrollpläne eingeschränkt.¹⁾ Die

1) Vgl. Minton, G., Inspection and Correction Error in Data Processing, in: Journal of the American Statistical Association, Vol. 64, No. 328, Menasha, Wisc. 1969, S.1256-1264.

ANNAHMEWAHRSCHEINLICHKEIT FÜR DIE SIGNIERUNG ALLGEMEINER MERKMALE IN DER ÜBUNGSPHASE



Quelle: United States Censuses of Population and Housing 1960, Quality Control of Preparatory Operations, Microfilming, and Coding, a.a.O., S.36.

Annahmekennlinien verschieben sich nach rechts. Das Konsumentenrisiko, also schlechte Arbeitsmassen als gut anzunehmen, steigt.

Wenn zurückgewiesene Arbeitsmassen zu 100 % kontrolliert und bereinigt werden, dann kann die durchschnittliche Auslieferungsqualität nach der Kontrolle einen bestimmten Höchstwert (größter Durchschlupf)¹⁾ nicht überschreiten. Bei sehr guter Qualität der zu kontrollierenden Arbeitsmassen entspricht deren Ausschußanteil etwa dem nach Durchführung der Annahmestichprobe, weil nur in wenigen Arbeitsmassen soviel schlechte Einheiten gefunden werden, daß eine vollständige Bearbeitung notwendig wird. Bei sehr schlechter Qualität der zu kontrollierenden Arbeitsmassen ist die durchschnittliche Auslieferungsqualität nach der Kontrolle wesentlich besser, weil viele Arbeitsmassen vollständig überprüft und bereinigt werden müssen. Das erhöht natürlich die Prüfkosten.

343 Hybride Verfahren

Diese Kombination aus Prozeßkontrollen und Annahmestichproben wurde in USA sehr erfolgreich eingesetzt. Sowohl zur Qualitätsregulierung der Arbeitsgänge als auch zur Qualitätsbeurteilung von Arbeitsmassen wurden spezifische Kontrollgrenzen fixiert. Wenn ein Arbeitsgang "außer statistischer Kontrolle" lag, bedeutete das im Normalfall nicht gleichzeitig die Rückweisung der Arbeitsmasse. Die Prozeßkontrollen wurden betont, weil auf diese Weise einzelne Mitarbeiter sehr schnell auf die mangelhafte Qualität ihrer Arbeitsleistung hingewiesen werden konnten und sich Rückweisungen von Arbeitsmassen mit der Konsequenz vollständiger Überarbeitung vermeiden ließen.

Der Qualitätsstandard wird durch die verfügbaren Ressourcen bestimmt. Da die Konsumenten, wie oben ausgeführt, bei Mehrzweck-

1) In der angelsächsischen Literatur "Average Outgoing Quality Limit" (AOQL).

statistiken wie Volkszählungen nicht in der Lage sind, einheitliche Qualitätsansprüche zu formulieren, ergibt sich für die Statistischen Ämter keine Konfliktsituation. Sie können im Prinzip jede durch wirtschaftliche Zwänge gebotene Qualität hinnehmen. Sie können lediglich darauf verweisen, daß sie verläßlichere Ergebnisse bei besserer finanzieller Ausstattung erreichen könnten. Bei gegebener Zielkonsistenz kann jedenfalls ein Stichprobenplan entwickelt werden, der bei einer bestimmten Trennschärfe die Prüfkosten minimiert. In USA wird ein differenziertes System Statistischer Qualitätskontrollen sowohl in der Phase der Datensammlung¹⁾ als auch in der Signierphase angewandt. Wegen der Abhängigkeit von den Gemeinden als Erhebungsstellen und von ehrenamtlich tätigen Zählern sind diese Kontrollen in der Bundesrepublik auf die Feldarbeit nicht übertragbar. Ihre Einführung zur Überwachung des Signierprozesses ist dagegen möglich und nach Meinung des Verfassers dringend erforderlich für eine spürbare Qualitätsverbesserung.

1) United States Censuses of Population and Housing 1960, Quality Control of the Field Enumeration, Hrsg. US Bureau of the Census, Washington, D.C. 1967.

35 Automatische Bestandskontrollen

Eine der wichtigsten Aufgaben einer Volkszählung besteht darin, die amtliche Bevölkerung in regionaler Gliederung festzustellen. Dazu genügt es keineswegs, die ausgefüllten Fragebögen "auszuzählen". Die Sicherung der Bestände im gesamten Arbeitsablauf der Aufbereitung stellt vielmehr bei der sehr großen Zahl statistischer Einheiten einer Volkszählung ein beträchtliches technisch-organisatorisches Problem mit Auswirkungen auf die Genauigkeit und die Aktualität der Ergebnisse dar. In einer Arbeit über die Fehlerbereinigung bei Volkszählungen kann auf diesen Aspekt nicht verzichtet werden.

351 Bevölkerungsbegriff in deutschen Volkszählungen

Drei Bevölkerungsbegriffe werden unterschieden:¹⁾

ortsanwesende Bevölkerung
 wohnberechtigte Bevölkerung
 Wohnbevölkerung

Zur Ermittlung der ortsanwesenden Bevölkerung wird jede Person in der Gemeinde gezählt, in der sie sich am Zählungstichtag aufhält. Dieses Konzept führte bei zunehmender Mobilität der Bevölkerung zu Ergebnissen, die stark von Zufälligkeiten bestimmt waren.²⁾

Seit 1925 wird unterschieden zwischen der wohnberechtigten Bevölkerung und der Wohnbevölkerung. Zur wohnberechtigten Bevölkerung gehören alle Personen, die am Erhebungsort einen Wohnsitz haben, auch wenn es nicht ihr Hauptwohnsitz ist. Dieses Konzept ist bedeutsam, weil die Kommunen (besonders z.B. Universitätsstädte) Infrastruktureinrichtungen für diese Gruppe planen müssen.

-
- 1) Nur einmal - bei der Volkszählung 1939 - fand die Bezeichnung "ständige Bevölkerung" Verwendung. Sie ergab sich aus der Wohnbevölkerung unter Absetzung der ihrer Dienstpflicht genügenden Soldaten und Arbeitsdienstangehörigen.
 - 2) Vgl. Ungern-Sternberg, R.v. und Schubnell, H., Grundriß der Bevölkerungswissenschaft (Demographie), Piscator, Stuttgart 1950, S.20/21.

Als amtliche Bevölkerung gilt jedoch nicht die wohnberechtigte Bevölkerung sondern die Wohnbevölkerung. Sie resultiert aus der Verminderung der wohnberechtigten Bevölkerung um die Personen mit weiterem Wohnsitz, die dort zugeordnet werden, von wo aus sie zur Arbeit oder Ausbildung gehen bzw. sich überwiegend aufhalten, wenn sie weder erwerbstätig sind noch in der Ausbildung stehen. Auf diesem Konzept beruhen viele Entscheidungen der öffentlichen Verwaltung wie des privaten Bereichs.

Die Bestandskontrollen haben die Aufgabe, die regionale Zuordnung der statistischen Einheiten zu sichern, darüber hinaus bei deutschen Volkszählungen, für Zwecke der Haushalts- und Familienstatistik die Haushaltsgröße zu bestimmen. Die Reduktion der wohnberechtigten Bevölkerung zur Wohnbevölkerung war bei der Volkszählung 1970 Inhalt eines Systems deterministischer Kontrollen und Korrekturen, das in Kapitel 6 beschrieben wird.

352 Ordnungssystem

Der sinnvolle Verbund von Ordnungsbegriffen wird hier als "Ordnungssystem" bezeichnet. Seine Ausgestaltung hängt von den organisatorischen Besonderheiten und den technischen Bedingungen des jeweiligen statistischen Projektes ab. Die Integration z.B. des Volkszählungsbogens 1970 in das Ordnungssystem wäre anders verlaufen, wenn es damals ein Lesegerät auf dem Markt gegeben hätte, das größere Belege als DIN A 4 hätte verarbeiten können; dann wäre als Erhebungsbogen eine Haushaltsliste (DIN A 3), wie sie bei früheren Volkszählungen benutzt wurde, vermutlich wieder zum Einsatz gekommen. So wurde der Erhebungsbogen als Individualbeleg konzipiert, der für jedes Haushaltsmitglied anzulegen war und das Problem der Sicherung des Haushaltszusammenhangs im Arbeitsablauf aufwarf.

Die Grafik auf der nächsten Seite zeigt das Ordnungssystem der Volkszählung 1970. Es besteht aus den Ordnungspapieren Zählerliste, Anstaltsliste, Gemeindeliste und enthält außerdem den Volkszählungsbogen. Zähler- und Anstaltslisten waren von den Zählern anzulegen, die Gemeindelisten waren von den Kommunen als Zählungsdienststellen auszufüllen.

Verbund der Zählpapiere der Volkszählung 1970

256 831

1-6

Bitte zur Abholung bereit-
halten ab: 1.06.70Arbeitsstättenzählung
27. Mai 1970

Arbeitsstättenbogen

KA 1

Vom Zähler auszufüllen!

Kreis: Detmold Gemeinde: Lage

Straße und Haus-Nr.: Hamelner Straße 1 Zahlbezirk-Nr. 0017

Bitte an den mit ○ gekennzeichneten Stellen die Erläuterungen auf Seite 3 und 4 beachten!

Ö. Kl. 8-17

U. Kl. 13-17

Name Kuhlmann Vorname Ernst

Gemeinde Lage Kreis Detmold

Straße/Haus-Nr. Hamelner Straße 1 Zahlbezirk 0017

Eigener Telefonanschluß in der Wohnung (Nr.) -

Bitte nur
Bleistift
benutzen

06 084 593

1 Geschlecht männlich ☒ weiblich ☐

2 Geburtsdatum Tag Monat Jahr

11 Sind Sie

erwerbstätig ☒ Landwirt ☐

mithelfend im Familienbetrieb ☐

arbeitslos/arbeitsuchend ☐

Zählerliste

(Volkszählung 1970)

Zählerlisten-Nr.: 112 311

Kreis DetmoldGemeinde LageStraße Hamelner Straße 1 - 17

Name des Haushaltsvorstandes bzw. Bezeichnung der Arbeits- stätte oder Anstalt	Hausnummer	Zahl der Pers. im Haushalt	Vz-Bogen-Nr.	Vz-Bogen-Nr.
01 <u>Kuhlmann, Ernst</u>	<u>001</u> <u> </u> <u> </u> <u>04</u>	1. Pers. <u>06</u> <u>084</u> <u>593</u>	2. <u> </u> <u> </u> <u>594</u>	<u> </u> <u> </u> <u> </u>
		3. <u> </u> <u> </u> <u>595</u>	4. <u> </u> <u> </u> <u>596</u>	<u> </u> <u> </u> <u> </u>
		5. <u> </u> <u> </u> <u> </u>		<u> </u> <u> </u> <u> </u>
02 <u>Konditorei Kohl</u>	<u>001</u> <u> </u> <u> </u> <u> </u>	1. Pers. <u> </u> <u> </u> <u> </u>	2. <u> </u> <u> </u> <u> </u>	<u>256831</u>
		3. <u> </u> <u> </u> <u> </u>	4. <u> </u> <u> </u> <u> </u>	<u> </u> <u> </u> <u> </u>
		5. <u> </u> <u> </u> <u> </u>		<u> </u> <u> </u> <u> </u>

Statistisches Landesamt
Nordrhein-Westfalen
Volkszählung 1970

Land Nordrhein-Westfalen

Reg.-Bez. DetmoldKreis DetmoldGemeinde LageGemeindeteil Ost

Gemeindeliste

05 7 33 114

Land RB Kreis Gemeinde

Nr. des Zähl- bezirks	Straße und Haus-Nr. (von bis)	Straßen- schlüssel	Nr. der		Zahl der Soldaten- bogen, die dem Zähler mitgegeben wurden	10%- Zählerliste (X)
			Zähler- liste	Anstalts- liste		
0017	<u>Hamelner Straße 1 - 17</u>	<u>0031</u>	<u>112311</u>			
0018	<u>Hamelner Straße 19</u>	<u>0031</u>		<u>091267</u>		

Statistisches Landesamt
Nordrhein-WestfalenAnstaltslisten-Nr.: 091 267

Name der Anstalt Henrietten-Stift Telefon 42 81

Straße Hamelner Straße Haus-Nr. 19

Gemeinde Lage Bitte zur Abholung bereit-
halten ab 1.06.70

Kreis Detmold

Volkszählung
27. Mai 1970

Liste für Anstalten, Heime u. ä. (stehe Verzeichnis auf der Innenseite)

Die Angaben im Kopf des Volkszählungsbogens waren von den Auskunftspersonen selbst zu machen. Die Kennziffern der Individualbögen waren in vorgegebene Haushaltsblöcke der Zählerlisten (je 5 Personen umfassend) bzw. Anstaltslisten einzutragen. In die Anstaltslisten waren darüber hinaus die Kennziffern der Fragebögen von Personen aufzunehmen, die keinen eigenen Haushalt führen. Die Ordnungspapiere enthielten die regionalen Ordnungsbegriffe wie Kreis, Gemeinde und Straße im Klartext sowie die Hausnummer. Die Kommunen hatten das Signieren der Regional- und Straßenschlüssel auf den Gemeindelisten vorzunehmen. Außerdem waren die Nummern der Zähler- und Anstaltslisten in die Gemeindelisten zu übernehmen.

Die automatischen Bestandskontrollen gliederten sich in zwei Phasen: In der ersten Phase wurde eine Paarigkeitskontrolle der Zähler- bzw. Anstaltslistennummern von den Belegen selbst mit denselben Nummern aus den Gemeindelisten heraus durchgeführt, zusätzlich wurden die manuell festgestellten und als Sollwerte vorgegebenen Haushalts- und Anstaltsgrößen mit der Summe der einzelnen abgelochten Fälle verglichen.¹⁾ Die Fehleranalyse dieser Phase zeigte Fehler aller vorangegangenen Arbeitsgänge:

Druckfehler der Zähler- bzw. Anstaltslistennummern, Übertragungsfehler durch Zähler und Zählungsdienststellen, Fehler der Vollzähligkeits- und Vollständigkeitskontrollen, Fehler der Loch- und Lochprüfarbeiten.

Die Beseitigung dieser Fehler war die Voraussetzung für die zweite Phase, in der eine Paarigkeitskontrolle der durch optische Beleglesung erfaßten Kennziffern der Individualbelege (Datei I) und der Kennziffern der Erhebungspapiere aus der Zählerliste (Datei II) erfolgte. Hier wurden auch die aus den Gemeindelisten abgelochten Regionalschlüssel mit einem fehlerfreien Leitband verglichen. Wenn durch Fehler in der manuellen

1) Deininger und Kerler verwenden dafür die Bezeichnung "Gruppenkontrolle auf Größenordnung"; Deininger, R., Kerler, E., Die Praxis automatischer Korrekturen und ihre Anwendung bei der Gebäude- und Wohnungszählung 1968, hektographierte Arbeitsunterlage, Hrsg. Statistisches Landesamt Baden-Württemberg, Stuttgart 1969, S.4.

Übertragung oder in der Ablochung der Regionalschlüssel nicht zutreffende aber gültige Begriffe entstanden, war die regionale Zuordnung von Personen gefährdet. Solche Fehler sind nicht auszuschließen, ihre Wahrscheinlichkeit hängt von der Kompaktheit der Codes - wie oben gesagt - ab. Die Entdeckung dieser Fehler muß späteren Arbeitsgängen vorbehalten bleiben.

Das Ziel der Bestandskontrollen war erreicht, wenn alle Datensätze mit den Informationen der Volkszählungsbögen (d.h. jede statistische Einheit) gesicherte Regionalschlüssel enthielten.

353 Paarigkeitskontrolle von Ordnungsbegriffen (Dateienvergleich)

Eine Paarigkeitskontrolle, wie sie im Rahmen der Bestandskontrollen der Volkszählung 1970 durchgeführt wurde, kann schematisch wie folgt dargestellt werden:

Datei I	Datei II
Kennziffer des Erhebungspapiers	Kennziffer des Erhebungspapiers aus dem Ordnungspapier
4010	4010
4011	4011
4012	4013
4016	4014
4017	4017
4019	4018
4020	4020

Die Kennziffern 4010, 4011, 4017 und 4020 werden "zusammengeführt", d.h. der Bestand wird gesichert. Die Kennziffern 4012, 4016 und 4019 erhalten den Fehlerhinweis "fehlt in Datei II", die Kennziffern 4013, 4014 und 4018 die Kennzeichnung "fehlt in Datei I". Das Fehlen der Kennziffern 4015 wird hierbei nicht reklamiert.

Die Beseitigung von Fehlern, die dadurch entstehen, daß z.B. zwei benachbarte Ziffern vertauscht sind, kann sehr zeitaufwendig sein, da die Sortierung der Kennziffern in aufsteigender Reihenfolge in den Dateien jeden Zusammenhang zerstört. Eine Auszählung des Statistischen Bundesamtes bei rd. 1,15 Mill. Personen hatte Unpaarigkeiten in 4 % der Fälle ergeben.

Paarigkeitskontrollen sind überflüssig, wenn die in den Kommunen in aufsteigender Nummernfolge verteilten Zählpapiere lückenlos zurückkommen. Diese Forderung kann bei Großzählungen nicht erfüllt werden, da die Nachforschung nach jedem Einzelbeleg unzumutbar ist.

354 Kleinräumige Gliederung

Totalzählungen ermöglichen den Nachweis von Ergebnissen für kleine regionale Einheiten. Für Zwecke der Stadtentwicklungs- und Regionalplanung sind grundstücksbezogene Daten am besten geeignet.¹⁾ Aus technischer Sicht hätte bei der Volkszählung 1970 - zumindest näherungsweise - die Möglichkeit bestanden, selbst so weitgehende Wünsche zu erfüllen, da Informationen über Straßenschlüssel und Hausnummern in den Ordnungspapieren enthalten waren. Ihre Kontrolle war in den Flächenstaaten der Bundesrepublik jedoch nicht möglich, da die Kontrollunterlagen nicht hinreichend zuverlässig waren und die Ablochung zu zeitaufwendig gewesen wäre. Wenn von den Kommunen Leitbänder mit gesicherten Regionalangaben zur Verfügung gestellt würden, könnten von der amtlichen Statistik Ergebnisse in kleinräumiger Gliederung als Standardtabellenprogramme aufbereitet werden und somit frühzeitig vorliegen. Eine andere Lösung bestünde darin, das regionale Schlüsselsystem um Informationen über Blöcke, Blockseiten oder auch Planquadrate zu erweitern. Unter Genauigkeitsaspekten sind allerdings ernste Zweifel angezeigt, Aggregate für 100, 50 oder gar 10 statistische Einheiten zu bilden. Leider fehlt bisher eine Fehlerforschung für kleine regionale Einheiten.

1) Dumas, L., Auswertungsmöglichkeiten des Zählungswerkes 1968/70 - Überlegungen unter dem Aspekt der Stadtentwicklung - in: Hamburg in Zahlen, Hrsg. Statistisches Landesamt Hamburg, Jg. 1972, März-Heft, S.78.

Bei der Volkszählung 1970 wurden kleinräumige Ergebnisse erst viele Monate nach den ersten Veröffentlichungen auf der Basis von Zählerlisten erstellt. Die Verzögerung war auf umfangreiche Rückfragen bei den Kommunen und Vorbereitung von Leitbändern zurückzuführen. Unter der Voraussetzung, daß die Kommunen eine sinnvolle Zuordnung von Straßen und Straßenabschnitten zu Zählerlisten vorgenommen hatten, stellte auch diese Lösung eine für Planer durchaus brauchbare kleinräumige Gliederung dar, wenngleich über die Genauigkeit hierbei ebensowenig ausgesagt werden kann.

In USA wird die Zuordnung von statistischen Einheiten zu kleinsten Räumen bereits vielfach automatisch vorgenommen (regional coding). Dort sind auf Magnetbändern regionale Verzeichnisse gespeichert, die Adressen in Verbindung mit numerischen Codes enthalten.

Gelegentlich werden regionale Begriffe auch schon akustisch eingegeben (soundex code) und die Verschlüsselung der Regionalangaben wie Bundesland, Verwaltungsbezirk, Stadt bzw. Gemeinde, Stadtteil, Straße usw. erfolgt maschinell. Dieses Verfahren verkürzt den Signier- und Kontrollaufwand erheblich und verbessert die Qualität der regionalen Zuordnung.

355 Verbesserung der Erfassungsgenauigkeit

Die Träger der amtlichen Statistik unterstellen, daß die von den Zählungsdienststellen abgelieferte Zahl von Erhebungspapieren eine verlässliche Grundlage für die Ermittlung der wohnberechtigten und damit auch der Wohnbevölkerung darstellt. Der Nachweis der auf den Zählungstichtag bezogenen "amtlichen Wohnbevölkerung" korrigiert die bis zu diesem Tag fortgeschriebene Wohnbevölkerung, die auf der natürlichen Bevölkerungsbe-
wegung (Geburten und Sterbefälle) und der Wanderungsbewegung (Zu- und Fortzüge) seit dem Stichtag der letzten Volkszählung beruht. So brachte die Gegenüberstellung der am 27. Mai 1970 ermittelten Wohnbevölkerung gegenüber der letzten Fortschreibung als Ergebnis im Bundesgebiet eine negative Differenz von rd. 859 000 Personen (1,4 %). In Nordrhein-Westfalen war sie ebenfalls negativ und betrug rd. 293 000 (1,7 %); bei männlichen

Personen 2,2 %, bei weiblichen Personen 1,2 %).

Im Interesse der Erfassungsgenauigkeit ist es wichtig, daß sich die amtliche Statistik zum frühestmöglichen Zeitpunkt die Information verschafft, in welchem Ausmaß solche Abweichungen zu erwarten sind.

Die automatischen Bestandskontrollen haben lediglich formalen Charakter. Sie sind im Sinne Weichselbergers endogene Kontrollen, da nur Daten aus den Unterlagen der jeweiligen Erhebung geprüft werden. Die Beseitigung von Erfassungsfehlern, d.h. Auslassungen bzw. Mehrfacherfassungen, wie sie im Normalfall schon auftreten, besonders aber Soldaten, Ausländer oder Personen im Anstaltsbereich betreffen, ist bei dieser Art von Kontrollen nicht möglich.

Der Verfasser schlägt daher vor, durch Einführung des Vergleichs mit externen Daten, den Kontrollen substantiellen Charakter zu verleihen, aus ihnen im Sinne Weichselbergers exogene Kontrollen zu machen. Diese könnten auf dem Vergleich der als Sollwerte gespeicherten Fortschreibungsergebnisse mit der maschinellen Summierung der in regionaler Ordnung auf die Datenträger gelangten Einzelbelege basieren.¹⁾ Der Bestand (wohnberechtigte Bevölkerung) könnte entweder durch Multiplikation mit einem empirisch ermittelten Koeffizienten (Wohnbevölkerung/wohnberechtigte Bevölkerung) oder durch eine frühzeitige maschinelle Bestimmung der Wohnbevölkerung (vgl. Kap. 6) bestimmt werden. Aufgrund von Erfahrungswerten über den Erfassungsfehler bei der laufenden Bevölkerungsstatistik könnten Kontrollgrenzen festgelegt werden, deren Überschreiten ein starkes Indiz gegen die vollständige Erfassung bei der Volkszählung aber auch gegen die korrekte Verarbeitung des Materials wäre.

1) Das Maschinenprogramm DIARY I des kanadischen Census 1971 sah unter anderem die Berechnung von Abweichungen zwischen Vorgabewerten und den kumulierten Fällen vor. Solche Vorgabewerte waren zählbezirksweise u.a. für die Summe der zu verarbeitenden Belege als auch für die Einwohner gespeichert. Die Vergleiche hatten demnach sowohl formalen als auch substantiellen Charakter; Final Report of the DIARY I Task Force: Detailed Specifications, hektographierte Arbeitsunterlage, Hrsg. Dominion Bureau of Statistics, Ottawa 1970, S.1-13.

Nacherhebungen in diesem Stadium wären noch relativ leicht durchzuführen. Wenn die Wohnbevölkerung später endgültig bestimmt ist und dann erst größere Abweichungen gegenüber der Fortschreibung sichtbar werden, sind Anpassungen weit- aus schwieriger.

Nacherhebungen dürfen den Arbeitsablauf nicht wesentlich verzögern. Das Problem einer sinnvollen Terminierung sowohl für diese Nachlieferungen als für die Anlieferung der Zähl- papiere generell müßte besser gelöst werden. Wenn den Zählungs- dienststellen keine ausreichende Zeitspanne für Erhebungs- kontrollen und Nacherhebungen gegeben wird, ist die Erfassungs- genauigkeit gefährdet. Da die ungeheueren Materialmengen durch die amtliche Statistik nur sukzessive verarbeitet werden können, müßte - etwa mittels der Netzplantechnik - eine re- gional differenzierte Terminplanung erfolgen.

Die schnellstmögliche Bekanntgabe vorläufiger Ergebnisse der Wohnbevölkerung (notfalls auch der wohnberechtigten Bevölke- rung) hat den Effekt eines Frühwarnsignals für die Zählungs- dienststellen. Sofern Konsumenten davon erfahren, müssen sie bereit sein, Entscheidungen unter Umständen zu revidieren, wenn die endgültigen Ergebnisse von den vorläufigen abweichen.

356 Erfahrungen

Die Aktualität der Volkszählungsergebnisse 1970 war beträcht- lich belastet durch Mängel der automatischen Bestandskontrollen und der damit verbundenen Korrekturen.¹⁾

In der zweiten Phase der automatischen Bestandskontrollen wurde besonders deutlich, wie schwierig die Integration des Lesegerätes in den Arbeitsablauf ist. Während die Ordnungsbe- griffe der Datei I durch Beleglesung sehr schnell auf die Magnet- bänder gelangten, war die Datei II infolge Ablochung erst sehr viel später verfügbar.

1) Es war geplant, 12 - 15 Monate nach der Erhebung Ergebnisse vorzulegen; vgl. Szameitat, K., Deininger, R., Die Vorbe- reitung des Zählungswerkes 1970, Teil 1, Das Gesamtkonzept, in: Wirtschaft und Statistik, H. 8, Wiesbaden 1967, S.452. Die Gemeindeblätter als erste Tabellen mit wichtigen Struktur- daten wurden 18 Monate nach dem Zählungsstichtag maschinell ausgedruckt.

Abhilfe könnte hier durch die Konzeption maschinell lesbarer Zählerlisten geschaffen werden. Um die Überwachung der stilisierten Ziffernschreibung sicherzustellen, müßte allerdings Personal aus der amtlichen Statistik diese Arbeit übernehmen.

Von grundlegender Bedeutung für den Erfolg eines überwiegend maschinell durchgeführten Projektes wie einer Volkszählung werden immer der Zeitbedarf und die Qualität der EDV-Programme sein. Wenn - wie in der amtlichen Statistik der Bundesrepublik - ein Programmierverbund der Statistischen Landesämter besteht, spielt dessen Effizienz gleichfalls eine wichtige Rolle. Verbesserungen sind hier sicherlich noch zu erzielen.

Automatische Bestandskontrollen stehen in Konkurrenz zu den manuellen Kontrollen der Ordnungspapiere auf Vollzähligkeit und Vollständigkeit. Unter Kostengesichtspunkten scheint auf den ersten Blick alles dafür zu sprechen, die manuellen Kontrollen vollständig durch die automatischen zu ersetzen. Dabei ist aber zu bedenken, daß der Qualitätsanspruch bei der regionalen Zuordnung der Bevölkerung absolut ist, d.h. jede statistische Einheit in der Gemeinde nachzuweisen ist, in der sie erfaßt wurde. Die Analysen und Fehlerbereinigungen im Anschluß an die automatischen Bestandskontrollen würden in sehr vielen Fällen zeitaufwendige Rückgriffe auf die Ordnungspapiere erforderlich machen, wenn durch manuelle Kontrollen nicht bereits ein erheblicher Fehleranteil beseitigt werden würde. Der Verfasser hält es jedoch für möglich, im Vorfeld der automatischen Bestandskontrollen die manuellen Kontrollen stichprobenhaft durchzuführen.

4 Fehlerallokation

Im Kontrollprozeß werden Einzelwerte und Wertekombinationen beanstandet und zur Korrektur zurückgewiesen. Um Korrekturverfahren anwenden zu können, die nur für die als falsch erkannten Werte Ersatzwerte einsetzen, werden Informationen über die Zuverlässigkeit einzelner Werte in Kombinationen benötigt. Fehlerallokation ist die Zuordnung von Zuverlässigkeitsmaßen zu Einzelwerten. Ansätze für die Konstruktion solcher Maße finden sich in der Literatur kaum.¹⁾ Die Vernachlässigung ist erstaunlich, da ohne die Bestimmung von Konfidenzgraden die Fehlerbeseitigung im Prinzip keine Basis hat. Bei Verzicht auf Fehlerallokation muß entweder das Risiko getragen werden, "gute" Informationen in Wertekombinationen zu vernichten, um der Kombination geschlossen zur Annahme zu verhelfen, oder es müssen im Falle der Einsetzung von nicht sofort in die Kombination passenden Werten Iterationen von Kontrollen und Korrekturen in Kauf genommen werden. Die in der Literatur vorgeschlagenen Verfahren der Fehlerallokation beruhen entweder auf der Anwendung des Bayes'schen Theorems oder auf der Ableitung von Zuverlässigkeitsmaßen in Abhängigkeit von der Zahl der Kontrollverletzungen. In USA waren bereits in den vierziger Jahren in Identitätsgleichungen je nach Zahl der Beanstandungen bei den Kontrolldurchläufen Fehlerwerte vergeben worden, die dann durch die Korrekturen beseitigt oder zumindest reduziert worden waren. Zuverlässigkeitsmaße für einzelne Variable können auch zu analytischen Zwecken im Rahmen der deskriptiven Fehlerforschung berechnet werden. Hier ist ihre operative Nutzung im Prozeß der Fehlerbereinigung gemeint.

41 Einzelfehler

Einzelfehler treten in einem bestimmten Datenfeld (d.h. bei den Werten einer Variablen) auf. Die relative Fehlerhäufigkeit -

1) Naus, J. I., Johnson, T. G., Montalvo, R., A Probabilistic Model for Identifying Errors in Data Editing, in: Journal of the American Statistical Association, Vol. 67, No. 340, Menasha, Wisc. 1972, S.943-950; Naus, J. I., Data Quality Control and Editing, a.a.O., S.123-149.

bezogen auf eine Variable - wird Einzelfehlerquote genannt. Wenn Naus sich der Bayes'schen Formel bedient,¹⁾ dann bezeichnet er diese Fehlerquote als "prior error rate", d.h. als "vorgegeben" bzw. "angenommen".

Informationen über diese Fehler können aus verschiedenen Quellen stammen. Es kann historisches Wissen über gleichartige oder ähnliche Erhebungen vorhanden sein, es können Rückschlüsse aus der laufenden Zählung gezogen werden (feedback) oder Erfahrungen aus allgemeiner Fehlerforschung vorliegen. Die Kenntnisse der Fehlerstrukturen älterer Projekte dürften die bedeutsamsten Grundlagen für die Bestimmung von Zuverlässigkeitsmaßen sein. Leider reicht dieses Wissen selten aus, da die Frageprogramme und die Erhebungsmethoden im Zeitablauf Veränderungen unterworfen sind. Ein generelles Problem ist der Grad der Fehleridentifikation in früheren Erhebungen. Je nach Intensität der Bemühungen und Wahl der Methoden können sich die Fehlerquoten unterscheiden und somit für den aktuellen Einsatz mehr oder weniger brauchbar sein. Immerhin läßt die Anwendung der Bayes'schen Formel Revisionen zu, wenn grobe historische Informationen durch bessere aus dem aktuellen Projekt ersetzt werden können.

Es können auch allgemeine Fehlererfahrungen genutzt werden. Empirische Untersuchungen haben gezeigt, daß bei zunehmender Zahl von Signiermöglichkeiten je Merkmal die Fehleranfälligkeit zunimmt. Der Verfasser hat andererseits beobachtet, daß bei starker Einschränkung der Signiermöglichkeiten (z.B. Verdichtung des Schlüsselpotentials der Wirtschaftssystematik auf 100 Codes) die Gefahr von Fehlzuordnungen in den Randzonen einzelner Branchen nicht unerheblich war. Gedächtniskontrollen bei Vermehrung der Stellenzahl sechs- und siebenstelliger Telefonnummern zeigten rapide Zunahmen der Fehlerquoten. Auch Kontrollen des Signierens dreistelliger alphabetischer bzw. numerischer Codes belegten, daß der größere Wertebereich des Alphabets das Auftreten von Fehlern begünstigte. Weiter ist bekannt, daß häufig vorkommende Codes noch zusätzlich dadurch überhöht werden, daß sie anstelle seltener vorkommenden eingesetzt werden. Das gilt z.B. für die Abgrenzung fremder Staatsangehörigkeiten wie auch für ungewöhnliche Religionsgemeinschaften.

1) Vgl. Fußnote S.73.

Einer praktischen Verwendung solcher allgemeinen Fehlerinformationen steht bisher entgegen, daß funktionale Zusammenhänge zwischen Fehlerquoten und ihren Ursachen nicht quantifiziert werden konnten. Das gilt auch für die Unterscheidung der Merkmale in "statische" und "dynamische". Verlässliche Informationen über die Stabilität der Variablen in Kombinationen wären ja gerade der Schlüssel für die Vergabe unterschiedlicher Zuverlässigkeitsmaße. Aber selbst als "statisch" erkannte - also ihrem Wesen nach besonders vertrauenswürdige - Merkmale können diese Eigenschaft verlieren, wenn sie auf dem Fragebogen schlecht platziert oder im Erhebungsvorgang nachlässig behandelt werden.

Nun liegt der Vorteil der hier erläuterten Methode darin, daß relativ grobe Annahmen über Einzelfehlerquoten nicht schädlich sind.

42 Unabhängigkeitsmodell

Diesem Modell liegt die Annahme zugrunde, daß das Auftreten von Fehlern bei einer Variablen unabhängig davon ist, ob Fehler in anderen Variablen gemacht werden. Das Modell setzt nicht voraus, daß die Variablen selbst unabhängig sind. Erforderlich ist, daß unabhängige Beobachtungen der Variablen vorliegen. Das Modell sieht die Verwendung von Einzelfehlerquoten nach der Einzelfeld-Kontrolle vor. Die bei dieser Kontrolle erkannten Fehler werden vom weiteren Ablauf ausgeschlossen. Es kommt also auf Fehler an, die in Kombinationskontrollen beanstandet werden. Bei Einzelfeld-Kontrollen ist eine Fehlerallokation nicht nötig; ein beanstandeter Wert gilt als "vollständig" unzuverlässig. Die Prämisse des Unabhängigkeitsmodells erlaubt es, Einzelfehlerquoten durch Multiplikation in beliebiger Weise zu Kombinationsfehlerquoten zusammenzusetzen. Es ist natürlich vorteilhaft, nur Einzelfehlerquoten ermitteln zu müssen und Kombinationsfehlerquoten nach Bedarf zu berechnen. Das Unabhängigkeitsmodell kann in Ausgangssituationen verwendet werden, in denen keine Informationen über spezifische Fehlerquoten verfügbar sind. Dann können gleiche Quoten unterstellt werden (z.B. 0,01 oder 0,001), wobei - wie sich weiter unten zeigt - die absolute Größenordnung keine Rolle

spielt. Naus stellt fest, daß die mit dem Unabhängigkeitsmodell erzielten Ergebnisse bei kleinen Fehlerquoten (ca. 0,01) gute Näherungen an die Ergebnisse anderer komplexer Modelle mit mehr Informationen über Abhängigkeitsbeziehungen darstellen würden. Das kann nicht verwundern, wenn die Multiplikation von Einzelfehlerquoten zu größenordnungsmäßig sehr kleinen Kombinationsfehlerquoten führt. Selbst in Fällen wie z.B. Datenfeldverschiebungen auf der Lochkarte, in denen die Anwendung des Modells offensichtlich ungeeignet ist, ließe sich weiter mit dem Modell arbeiten, wenn die betreffenden Lochkarten aussortiert würden. Nun gelten nicht nur Datenfeldverschiebungen auf Lochkarten als Beispiele für die Untauglichkeit des Modells, vor allem Mängel des Fragebogens und der Beantwortung der Fragen schlechthin führen zu Abhängigkeiten der Informationen und damit der Fehlerstruktur. Aber zur Berechnung des Zuverlässigkeitsmaßes kann die Datengrundlage natürlich auch aus realistischeren Modellen herrühren.

43 Zuverlässigkeitsmaß

Nach der für die Anwendung der Bayes'schen Formel üblichen Diktion wird das Resultat des Algorithmus als "posteriore Fehlerwahrscheinlichkeit" bezeichnet und als Zuverlässigkeitsmaß interpretiert. Dieses Maß hängt nun außer von den "vorgegebenen" Fehlerquoten von den Kontrollverletzungen durch Merkmalskombinationen ab. Wenn z.B. in einer bivariaten Kombination eine Variable aus einem geprüften Leitband stammt, d.h. als sehr zuverlässig gelten kann, dann aber eine Reihe von Kontrollen verletzt, in denen sie vorkommt, muß die intuitiv als hoch empfundene Zuverlässigkeit durch Verknüpfung mit der anderen Information revidiert werden.

Allgemeine Lösung:¹⁾

$(X_1, X_2, \dots, X_n)$: gemeinsam verteilte Zufallsvariable

$(x_1, x_2, \dots, x_n)$: bestimmte Realisation der Zufallsvariablen

A_i : Ereignis, daß x_i falsch ist

$R_{12\dots n}$: Ereignis, daß die Realisationen der n Variablen in den Annahmebereich fallen

1) Vgl. Naus, J. I., Johnson, T. G., Montalvo, R., A Probabilistic Model for Identifying Errors in Data Editing, a.a.O., S.945.

Eine Kontrolle ist die Überprüfung dieses Ereignisses. Eine deterministische Kontrolle ist dann wie folgt definiert:

$$P \left(\bigcup_{i=1}^n A_i \mid \bar{R}_{12\dots n} \right) = 1$$

Das heißt, wenn die Überprüfung des genannten Ereignisses zu Beanstandungen führt, dann muß mindestens der Wert einer Zufallsvariablen falsch sein. So ist z.B. $P(A_1 \mid \bar{R}_{12})$ die bedingte Wahrscheinlichkeit, daß der Wert der Variablen 1 falsch ist, wenn eine bivariate Kombinationskontrolle zwischen den Variablen 1 und 2 verletzt ist. Umfangreiche empirische Informationen über die Fehleranfälligkeit der verknüpften Variablen wären notwendig, um diese Wahrscheinlichkeiten zu schätzen. Daher führen die Autoren anstelle dieser Wahrscheinlichkeiten solche ein, die ohne das Wissen über die Eigenart und die gemeinsamen Verteilungen der verknüpften Variablen auskommen. $P(A_1 \mid \bar{R}_{12})$ wird ersetzt durch $P(A_1 \mid A_1 \cup A_2)$. $A_1 \cup A_2$ ist eine logische Konsequenz von $\bar{R}_{12}$. Umgekehrt gilt das nicht unbedingt, da nicht jeder Fehler zu einer Kontrollverletzung führt. Die Autoren weisen nach, daß die ersatzweise eingeführten Wahrscheinlichkeiten - besonders bei kleinen "vorgegebenen" Einzelfehlerquoten - zu guten Näherungen führen.¹⁾ Wenn $P(A_1) = p_1$, $P(A_2) = p_2$, zudem die Ereignisse A_1 und A_2 unabhängig sind, dann ergibt sich mittels der Bayes'schen Formel die Wahrscheinlichkeit dafür, daß x_1 falsch ist, mit

$$P(A_1 \mid A_1 \cup A_2) = p_1 / (p_1 + p_2 - p_{12})$$

Dazu ein Zahlenbeispiel:

Variable 1 habe nach dem Durchlaufen der Einzelfeld-Kontrolle eine bereinigte Fehlerquote von 0,03, Variable 2 habe eine solche Fehlerquote von 0,01. Beide Variable seien kombiniert und verletzten Kontrollen, in denen sie vorkommen. Das Maß für die posteriore Wahrscheinlichkeit dafür, daß Variable 1 falsch ist,

1) Naus, J. I., Johnson, T. G., Montalvo, R., A Probabilistic Model for Identifying Errors in Data Editing, a.a.O., S.945 und S.949.

lautet dann:

$$p_1 / (p_1 + p_2 - p_{12}) = 0,03 / 0,0397 \doteq 0,75$$

Die Wahrscheinlichkeit, daß ein Wert der Variablen 1 falsch ist, beträgt somit etwa 0,75, während sie für einen Wert der Variablen 2 etwa 0,25 ausmacht. Bei kleinen "vorgegebenen" Einzelfehlerquoten ist die Kombinationsfehlerquote vernachlässigbar klein (hier: 0,0003), das Ergebnis liegt also nahe bei

$$p_1 / (p_1 + p_2)$$

Das Unabhängigkeitsmodell mit kleinen Fehlerquoten zeigt demnach für den bivariaten Fall, daß die posterioren Wahrscheinlichkeitsmaße im selben Verhältnis stehen wie die "vorgegebenen".

Wenn z.B. zwei bivariate Kontrollen (Verknüpfung der Variablen 1 und 2 sowie der Variablen 2 und 3) gleichzeitig verletzt werden, dann kann unterstellt werden, daß ein Fehler in der Variablen 2 gemacht wurde, der zu beiden Kontrollverletzungen führte. Da die meisten Fehler als Einzelfehler vorkommen, wird das als plausibelste Erklärung für die Beanstandung einer Datenmenge angesehen. Die Formel, die das Zuverlässigkeitsmaß für Variable 2 in diesem Fall bestimmte, würde lauten:

$$p_2 / (p_2 + p_{13} - p_{123})$$

Bei kleinen "vorgegebenen" Fehlerquoten reduziert sich der Begriff auf

$$p_2 / p_2 \doteq 1$$

Danach ist fast sicher, daß die Variable 2 "vollständig" unzuverlässig ist.

Dieses Verfahren der Fehlerallokation erweckt den Anschein, daß es zumindest in Fällen, in denen keine brauchbaren "vorgegebenen" Fehlerquoten existieren, auf Intuition beruht, die durch den Formalismus gestützt wird. Andererseits scheinen bereits verlässliche, als Vorinformationen eingebrachte Fehlerquoten, in ihrer Verlässlichkeit lediglich bestätigt zu werden. In einem anderen

Ansatz werden die Kontrollverletzungen isoliert untersucht, wobei sich ebenfalls Zuverlässigkeitsmaße bestimmen lassen.

44 Ableitung von "Erklärungsmaßen" aus deterministischen Kontrollen

Wenn beim Durchlaufen eines Wertes durch verschiedene Kontrollen, in denen die Variable enthalten ist, Beanstandungen festgestellt werden, erscheint die Annahme logisch, daß dieser Wert für die Beanstandungen verantwortlich sei. Im Prozeß der Fehlerbereinigung können diese verletzten Kontrollen aufgelistet werden und gleichzeitig "logische Erklärungen" dafür. Es stellt sich die Frage, wie groß die kleinste Erklärungsmaße von Variablen ist, mit der die Gesamtheit der verletzten Kontrollen vollständig erklärt werden kann. Hier wird der Zusammenhang zu der von Fellegi/Holt erarbeiteten Lösung deutlich, die ebenfalls die kleinste zu verändernde Anzahl von Datenfeldern je Datensatz bestimmen unter der Bedingung, daß ihre sinnvolle Veränderung keine Kontrollverletzungen mehr hervorruft. Naus nimmt bedauerlicherweise keine Stellung zur anschließenden Fehlerbeseitigung.

Der Grundgedanke soll an einem Beispiel erläutert werden:

Kontrolle	Kontrollrahmen: Verknüpfung der Variablen
1	1,2,3,4
2	1,2,5,6
3	1,3,5,7
4	2,3,6,7
5	4,6,7
6	4,8,9
7	1,5,8

In den sieben Kontrollen kommen 9 Variable vor. Zunächst ist nur bekannt, daß diese Variablen in ihrer Kombination zu Beanstan-

dungen des Datenmaterials geführt haben. Vorausgesetzt, das Verzeichnis der deterministischen Kontrollen ist vollständig, dann erklären die Variablen 1, 2, 4 - wenn sie falsch sind - die gleichzeitige Verletzung der sieben Kontrollen. Das gleiche gilt aber auch z.B. für die Variablen 1, 3, 4. Beides sind demnach logische Erklärungsmengen und gleichzeitig die kleinsten Mengen. Insgesamt lassen sich 14 Mengen bestimmen, mit denen die Kontrollverletzungen zu erklären sind. Wenn die Fehlerquoten klein sind und die Fehler unabhängig voneinander auftreten, dann ist die kleinere Anzahl miteinander kombinierter Variablen die wahrscheinlichere, in diesem Fall also drei. Naus meint nun, daß der Analytiker für diese kleinsten Erklärungsmengen die Werte findet, die gleichzeitig falsch sind. Wie das zu tun wäre, dafür gibt er keine Hinweise. Er erläutert lediglich die Techniken, mit denen die Erklärungsmengen grundsätzlich abzuleiten sind und wie die kleinsten Mengen bestimmt werden. In der Praxis sind sicherlich auch Fälle denkbar, in denen Variable fehlerhaft sind, ohne daß Kontrollen verletzt werden.

Aus diesem System können auch Fehlerwahrscheinlichkeiten berechnet werden, wenn die "vorgegebenen" Fehlerquoten bekannt sind und das Unabhängigkeitsmodell gilt. Für jede Merkmalskombination der Erklärungsmengen kann die Kombinationsfehlerquote ermittelt werden. Alle Kombinationsfehlerquoten addiert, die eine bestimmte Variable beinhalten, und auf die Summe der Kombinationsfehlerquoten bezogen, ergibt das gewünschte Maß. Eine gute Annäherung ist festzustellen, wenn die Häufigkeit des Vorkommens einer Variablen in den Erklärungsmengen auf die Gesamtzahl der Erklärungsmengen bezogen wird. Variable 1 z.B. kommt siebenmal vor in 14 Erklärungsmengen, d.h. sie erhält ein Zuverlässigkeitsmaß von 0,5 zugeordnet. Die Bestimmung aus den Kombinationsfehlerquoten ergibt einen Wert von 0,56. Naus läßt offen, ob seine Erkenntnisse operativen Zwecken dienen sollen oder - "wenn keine Möglichkeit zur Fehlerbereinigung gegeben ist"¹⁾ - der Analytiker immerhin eine Vorstellung von den Fehlern in den Daten und wahrscheinliche Erklärungen für Kontroll-

1) Naus, J. I., Data Quality Control and Editing, a.a.O., S.157.

verletzungen hat. Im Gegensatz zu der Studie von Fellegi/Holt wird hier auch nicht der Versuch gemacht, die für das System benötigte Vollständigkeit der Kontrollen zu erreichen. Vor allem die Tatsache, daß sich keine Überlegungen zur Korrekturphase anschließen, lassen die Arbeit als einigermaßen "künstlich" erscheinen. Die sehr große Zahl der bei Volkszählungen zur Anwendung kommenden Kontrollen und die riesigen Datenmengen, die mit Sicherheit zur Verletzung jeder einzelnen Kontrolle führen, entziehen diesem Ansatz den Boden, machen es andererseits aber auch sehr zweifelhaft, ob die - wenn auch besser durchdachte - Lösung von Fellegi/Holt praktisch relevant ist.

5 Fehlerbeseitigung

51 Bestimmung wichtiger Begriffe

Unter Fehlerbeseitigung wird der Korrekturprozeß verstanden, der sich an die Kontrollen - und an die möglicherweise durchgeführte Fehlerallokation - anschließt.

Dieses Kapitel umfaßt auch die Parameterschätzung von beobachteten fehlerbehafteten Verteilungen. Die Fehlerbeseitigung liegt hierbei in der Löschung der beanstandeten Werte in den Datensätzen. Somit besteht kein Unterschied zwischen einem Blank als teilweisem Informationsausfall und einem Blank, das durch Löschen einer Information entstanden ist.

Die Einzelfall-Korrektur hat dagegen die sinnvolle Substitution zurückgewiesener Werte oder Wertekombinationen bzw. im Falle von Blanks die Vornahme sinnvoller Einsetzungen zum Inhalt.

Einzelfall-Korrekturen werden in stochastische und deterministische eingeteilt. Stochastisch sind die Korrekturen dann, wenn die benötigten Informationen einer Zufallsauswahl entstammen. Deterministisch sind die Korrekturen, wenn aus einer gegebenen Datenkonstellation heraus die Substitution oder Einsetzung "zwangsläufig" erfolgt. In Analogie zu den Kontrollen läßt sich auch hier von einem "Wenn-Dann"-Schema sprechen: Wenn das Alter "unter 16 Jahre" und das Geschlecht "weiblich" ist, dann wird (zumindest für Deutsche) die beim Familienstand etwa vorhandene Ausprägung "verheiratet" in "ledig" umgesetzt.

Korrekturen werfen drei Probleme auf: Die Fehlerallokation, also die Fehlerbestimmung einzelner Werte in Wertekombinationen, das Abschneiden von Häufigkeitsverteilungen, also die Beschränkung der Korrekturen auf bestimmte Intervalle der Verteilungen, und die Iteration. Sie tritt bei stochastischen Korrekturen auf, wenn sich ein korrigierter Einzelwert in einer Wertekombination als inkonsistent erweist. Die Wiederholung einzelner Kontrollen oder ganzer Kontrollabschnitte mit dem Ziel, widerspruchsfreie Daten zu erhalten, wird Iteration genannt.

52 Parameterschätzung

Abweichend von der üblichen Diktion wird hier unter Parameterschätzung die Schätzung von Lage- und Streuungsparametern der unbekannten Gesamtheit sowie die Schätzung von Totalen (Aggregaten) verstanden. Bei der Fehlerbereinigung bedeutet das den Verzicht auf die Korrektur einzelner fehlerhafter Datensätze. Die Schätzung beruht stattdessen auf den vorhandenen Informationen in den Datensätzen oder auf der Kenntnis spezifischer Fehlerquellen. Das Verfahren bietet sich besonders im Non Response-Fall an, d.h. sowohl bei totalen Informationsausfällen kompletter Datensätze wie bei teilweisen Informationsausfällen einzelner Datenfelder. Die Fehleridentifikation braucht hier, zumindest für den teilweisen Informationsausfall, nicht besonders aufwendig gestaltet zu sein. Im Prinzip genügen Einzelfeld-Kontrollen. Um allerdings auch Rückweisungen wie teilweise Informationsausfälle zu behandeln, verlangt die Fehleridentifikation einen schwierigeren Ansatz.

Wenn in Stichprobenerhebungen auf alle beanstandeten Datensätze verzichtet wird, also nur die unbeanstandeten Datensätze für die Schätzung der Gesamtheitstotalen zugrunde gelegt werden, dann kann die Anpassung des Hochrechnungsfaktors vom reziproken Wert des Stichprobenauswahlsatzes an einen höheren Wert nach Maßgabe der fehlerfreien Datensätze vorgenommen werden. Ein solches Hochrechnungsverfahren ist auch für Totalzählungen denkbar unter der Annahme, die Gesamtheitstotale seien zufällig zustandegekommene Werte.

521 Schätzung von Antwortwahrscheinlichkeiten

In die Hochrechnung lassen sich bei der Behandlung von Non Response-Fällen empirische Erfahrungen mit Auskunftspersonen, die zum Zählungszeitpunkt nicht anzutreffen waren oder generell schwer erfaßbar sind ("not at home"-respondents), einbeziehen. In einem kürzlich erschienenen Diskussionsbeitrag¹⁾ werden Verfahren der Fehlerbereinigung solcher Fälle, die auf vorhandenen Informationen "ähnlicher" Auskunftspersonen beruhen, kritisiert.

1) Nargundkar, M. S., Joshi, G. B., Non Response in Sample Surveys, Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975.

Es wird eine Lösung vorgeschlagen, die über die Schätzung der Gesamtheitstotalen hinaus das schwerwiegendere Problem der Schätzung der Stichproben-Varianzen beinhaltet. Es liegt die Modellvorstellung zugrunde, daß der totale wie der teilweise Informationsausfall ein Zufallseignis ist, dessen Wahrscheinlichkeit (response probability) sich ermitteln ließe. Die Wahrscheinlichkeit, daß eine Einheit als Informationsträger ausfällt (u_i) bzw. die Wahrscheinlichkeit, daß ein Merkmal einer Einheit nicht berichtet wird (w_{ik}), wird zusammen mit der Auswahlwahrscheinlichkeit (π_i) verwendet bei der Schätzung der Gesamtheitstotalen und der Stichproben-Varianzen.

Für die Schätzung der Gesamtheitstotalen schlagen die Autoren die Horvitz-Thompson-Schätzfunktion vor, d.h. ein Verfahren, das Zufallsziehungen aus einer endlichen Gesamtheit (ohne Zurücklegen) - also mit ungleichen Auswahlchancen - vorsieht.¹⁾ Als Varianz wird die Yates-Grundy-Varianz benutzt.

Zur Schätzung der Gesamtheitstotalen und der Stichproben-Varianzen werden nur die u_i bzw. w_{ik} der Einheiten, die Informationen gebracht haben, benötigt. So könnten nach Auffassung der Autoren - ohne daß a priori-Informationen über die Gesamtheit vorliegen - besondere Informationen über die zufällig ausgewählten Auskunftspersonen in der Phase der Datensammlung gewonnen werden, um die erforderlichen Antwortwahrscheinlichkeiten empirisch schätzen zu können. Die Autoren halten es weder für praktikabel noch für notwendig, die Antwortwahrscheinlichkeiten in jedem Einzelfall zu schätzen. Es würde genügen, die Gesamtheit in homogene Schichten nach dem Kriterium der Antwortfähigkeit einzuteilen, wobei alle Individuen einer Schicht übereinstimmende Antwortwahrscheinlichkeiten zugeordnet erhalten sollen.

Es dürfte eine kaum lösbare Aufgabe darstellen, die wichtigsten Einflußfaktoren für die Auskunftserteilung - Zeit und Ort der Befragung, Schwierigkeitsgrad der Fragestellung, Fragebogengestaltung, Einstellung der Befragten und der Interviewer - zu quantifizieren. Hinzu kommt, daß durchaus nicht sicher ist, ob Informationsausfälle Zufallsergebnisse sind.

1) Vgl. Kendall, M. G., Buckland, W. R., Stichwort "Horvitz & Thompson Estimator", in: A Dictionary of Statistical Terms, 3. Aufl., Oliver & Boyd, Edinburgh 1972, S.68.

522 Maximum Likelihood-Schätzungen

In einer Arbeit von Hartley/Hocking wird eine ausführliche Beschreibung des Problems unvollständiger Daten mit zahlreichen Literaturhinweisen gegeben.¹⁾ Es werden einheitliche Analyse-Methoden vorgeschlagen, um - wie die Autoren meinen - den inzwischen entstandenen "Wildwuchs" zu lichten. Im Mittelpunkt stehen solche Verfahren, die Erweiterungen der Analyse vollständiger Daten darstellen, weil nach Auffassung der Autoren das Problem nur so in seiner allgemeinsten Form behandelt werden kann. Das Schätzprinzip ist entweder die Maximum Likelihood-Methode oder basiert zumindest darauf.

Eine Einschränkung wird insofern vorgenommen, als das Problem auf die Schätzung von Parametern aus beobachteten Daten beschränkt wird. Ausgangspunkt ist die Likelihoodfunktion

$$L = L(Y, X, \Theta)$$

wobei

Y eine Matrix statistischer Variabler,
X eine Matrix bekannter "Konstanten",
 Θ ein Vektor unbekannter Parameter ist.

Die Aufgabe der Datenanalyse besteht nun in der Schätzung der Θ aus den beobachteten Variablen Y und den gegebenen Konstanten X.

Die Autoren stellen die möglichen Aussagen über die unvollständigen Daten in folgender Übersicht zusammen:

- Y : Die Informationen über einzelne Elemente der Variablen Y sind unvollständig.
- X : Die Informationen über einzelne Elemente der Konstanten X sind unvollständig.
- m : Die Unvollständigkeit beruht auf dem Fehlen von Elementen.
- g : Von den fehlenden Elementen ist bekannt, daß sie in bestimmten Grenzen liegen.
- c : Die Elemente sind stetige Variable bzw. Konstante.
- d : Die Elemente von Y sind diskrete Variable bzw. die Elemente von X sind diskrete Konstante.

1) Hartley, H. O., Hocking, R. R., The Analysis of Incomplete Data, in: Biometrics, Vol. 27, Hrsg. Biometric Society, Raleigh, NC. 1971 S.783-823.

Die einzelnen Komponenten der Übersicht können in unterschiedlicher Konstellation auftreten. Die Autoren unterscheiden acht verschiedene Fälle. Sie bezeichnen selbst zwei davon als die wichtigsten. Das ist zuerst die Konstellation (Y, m, c) . Dabei soll es sich um die Schätzung der Parameter in multivariaten Normalverteilungen handeln. Die Variablen sind hier p -dimensionale Normalvektoren Y , von denen angenommen wird, sie seien als n unabhängige Stichproben aus $N(\mu, \Sigma)$ gezogen worden. Der Mittelwert-Vektor μ und die $p \times p$ -Kovarianz-Matrix Σ müssen geschätzt werden, wenn einzelne Elemente der Y -Vektoren fehlen.

Die zweitwichtigste Konstellation ist (X, m, c) , genauer, die lineare Regressions-Analyse mit fehlenden "fixierten" X -Werten. In diesem Fall liegt das Modell

$$Y = X\theta + e$$

vor, wobei die e unabhängige Werte aus $N(0, \sigma^2)$ sind und bestimmte Elemente von X fehlen. Sowohl θ als auch die fehlenden x -Elemente sind zu schätzen.

Der Verfasser sieht die Regressions-Schätzung in der vorliegenden Arbeit als Einzelfall-Korrektur, bei der zuerst die fehlenden x -Werte in den Datensätzen geschätzt und erst anschließend - quasi als "Qualitätskontrolle" - die Parameter der Kovarianz-Matrix betrachtet werden. Diese Lösung wird weiter unten dargestellt.¹⁾

Von der historischen Entwicklung her gesehen, haben Wilks (1932) und Rao (1956) zuerst die Parameter einer bivariaten normalen Gesamtheit mit fehlenden Werten geschätzt. Matthai (1951) hat die multivariate normale Gesamtheit untersucht und Lösungen für den trivariaten Fall angegeben.²⁾

Die Methode der genannten Autoren ist die simultane Schätzung der Parameter der bi- oder trivariaten normalen Gesamtheit als Maximum Likelihood-Lösung. Es handelte sich noch nicht um geschlossene Lösungen.

1) Vgl. Abschnitt 53 "Einzelfall-Korrekturen" S.92 ff.

2) Vgl. Buck, S. F., A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer, in: Journal of the Royal Statistical Society, Ser. B, Vol. 22, London 1960, S.302/303.

Edgett (1956) fand die Maximum Likelihood-Lösungen in expliziter Form für die Parameter einer trivariaten normalen Gesamtheit für den Spezialfall, daß einige Stichprobenbeobachtungen für nur eine der Variablen fehlten.

Anderson, Nicholson (beide 1957) haben dasselbe Problem betrachtet und die beste mathematische Methode bei der Berechnung der Maximum Likelihood-Schätzungen der Parameter aus unvollständigen multivariaten Normalverteilungen aufgezeigt.

Für mehr als zwei Variable ist die analytische Methode recht kompliziert. Deswegen setzt man Verfahren der numerischen Mathematik (meist Iterationsverfahren) ein, die rechenaufwendig sind und heute meist mit Hilfe der EDV durchgeführt werden.

Das gilt z.B. für die Untersuchung von Hocking, Huddleston, Hunt.¹⁾ Um es zu wiederholen, Ziel ist die Schätzung von Grundgesamtheitsparametern (Mittelwert, Varianz, Kovarianz) direkt aus den vorhandenen Daten, ohne daß zuerst Werte für die fehlenden Daten in einzelne Datensätze eingesetzt werden. Wichtige Prämissen sind, daß die Daten multivariat normalverteilt sind und daß die Werte in den besetzten Datenfeldern fehlerfrei sind, d.h. die Kovarianz-Matrix, die aus den kompletten Datensätzen geschätzt wird, unverzerrt ist.

Gegeben seien N Datensätze (Vektoren), von denen jeder eine Beobachtung aus der p -dimensionalen normalverteilten Gesamtheit $N(\mu, \Sigma)$ darstellt. Sie werden in T Gruppen (Schichten) eingeteilt. Schichtungsmerkmal ist die Unvollständigkeitsstruktur der Datensätze (Blanks in einzelnen Datenfeldern bzw. in mehreren gleichzeitig, wobei Linearkombinationen - z.B. Summen - einbezogen sind). Gruppenweise werden für die vorhandenen Beobachtungen der Mittelwert-Vektor und die Kovarianz-Matrix berechnet. Durch eine Indikator-Matrix wird die Unvollständigkeitsstruktur in jeder Gruppe beschrieben:

$$Y_{ti} = D_t \cdot Z_{ti}; \quad \begin{array}{l} t = 1, 2, \dots, T \\ i = 1, 2, \dots, n_t \end{array}$$

Y_{ti} : Vektor der Länge q_t der zu untersuchenden Einzelwerte (Datenfelder) bzw. von Linearkombinationen; q_t ist deren Anzahl in Gruppe t .

1) Hocking, R. R., Huddleston, H. F., Hunt, H. H., A Procedure for Editing Survey Data, in: Journal of the Royal Statistical Society Applied Statistics, Ser. C, Vol.23, No.2, London 1974, S.121-133.

D_t : Indikator-Matrix ($q_t \times p$ -Matrix)

Z_{ti} : vollständiger p -dimensionaler beobachteter Datenvektor
in Gruppe t

Bei $p = 3$ (d.h. der Datensatz besteht aus drei Datenfeldern) und unter der Annahme, in Gruppe 1 fehlten Informationen im zweiten Datenfeld, lautet die Indikator-Matrix

$$D_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Sei $p = 3$ und unter der Annahme, in Gruppe 2 werde nur das erste Datenfeld sowie die Summe der beiden übrigen untersucht, lautet die Indikator-Matrix

$$D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Die D_t transformieren die Z_{ti} derart, daß die Y_{ti} aus den beobachteten Vektoren Z_{ti} durch Selektion bzw. Linearkombination einzelner Komponenten entstehen. Die Y_{ti} sind daher auch multivariat normalverteilt.

Von den Autoren wird nun zur numerischen Konkretisierung ein eigens entwickeltes Programm (ESTMAT) eingesetzt. Dem Rechenprogramm als Eingaben dienen die Indikator-Matrix, die Mittelwert-Vektoren und die Kovarianz-Matrizen. Das Programm verknüpft die Informationen und nimmt Maximum Likelihood-Schätzungen des alle Gruppen umfassenden Mittelwert-Vektors und der Kovarianz-Matrix unter der Voraussetzung der Normalität vor. Mit Hilfe dieser Ergebnisse können nun Kontrollen auf Extremwerte und - wenn das gewünscht wird - Einzelfall-Korrekturen durch Verwendung der multiplen Regressionsrechnung durchgeführt werden.

Die Autoren verwenden zwei Beispiele aus der Landwirtschaft, eine Erhebung der Milchproduktion in Wisconsin im Jahre 1971 und eine Viehzählung in Texas aus dem Jahre 1969. In beiden Fällen wurde versucht, die Daten zunächst durch eine Quadratwurzel-Transformation der Normalverteilung anzunähern.

Die Ergebnisse von ESTMAT werden dann mit den Ergebnissen der Gesamtheit vollständiger Datensätze (in die die "Weglassungen" zufällig hineingebracht worden waren) und mit denen des Pooling, d.i. die Parameterschätzung aller vorhandenen Informationen zu

einem Merkmal, verglichen. Bei der Analyse der Kovarianz-Matrix wird von den Autoren - wie das in vielen Fällen üblich ist - die Kovarianz-Matrix herangezogen, die aus den nach der "Fehlererzeugung" noch verbliebenen vollständigen Datensätzen stammt.

Die Hauptergebnisse zeigen die geschätzten Varianzen der Schätzwerte von μ und die geschätzten Varianzen der Schätzwerte der Diagonalen σ_{ii} der Kovarianz-Matrix Σ . Es kommen die Genauigkeitsverluste aufgrund der fehlenden Daten sowie die relative Vorteilhaftigkeit von ESTMAT in Vergleich zu den anderen Verfahren zum Ausdruck. Um die Robustheit des Verfahrens gegen Abweichungen von der Normalität zu testen, werden die wichtigsten Ergebnisse auch für die Daten ohne vorherige Quadratwurzel-Transformation angewandt. Das Verfahren erweist sich auch bei leichter Abweichung von der Normalform als robust.

In einem Fall, in dem ein Datenfeld die Summe der übrigen enthält (Identitätsgleichung), wird gezeigt, daß der Varianzanstieg der Schätzwerte für μ und σ_{ii} geringer ist, wenn die Summe in der Eingabe berücksichtigt wird, als bei Weglassen dieser Information.

Natürlich reichen zwei Beispiele nicht aus, um das Verfahren ESTMAT endgültig würdigen zu können. Unterschiedliche Stichproben-Variabilität und die Korrelationsstruktur sind jeweils in Rechnung zu stellen. Es kann keine Garantie für die Konvergenz des Verfahrens und - auch wenn es konvergiert - keine Sicherheit für das Maximum der Likelihood-Funktion gegeben werden. In der Anwendungspraxis der Autoren wurde relativ schnell Konvergenz erzielt, oft bereits nach fünf Iterationen. Als sicher gilt die Maximierung der Likelihood-Funktion nur im Falle von zwei Gruppen. Wenn Konvergenz erreicht ist, können die schließlich berechneten μ und σ_{ii} zur Einsetzung von Werten für fehlende Angaben bei Einzelfällen unter Verwendung geeigneter Regressionsgleichungen verwendet werden.

Abgesehen von der strengen Voraussetzung der Normalität, ist die Voraussetzung, daß die als Anfangswerte fungierenden vollständigen Datensätze fehlerfrei sind, unrealistisch. Ein Verbesserungsvorschlag wäre, die Kovarianz-Matrix weiterhin unter der Prämisse der Fehlerfreiheit der besetzten Felder zu schätzen, anschließend aber Kontrollen durchzuführen mit größeren kritischen Werten als bei den

unverzerrten Schätzungen der Kovarianz-Matrix. Die beanstandeten Extremwerte fielen dann heraus und die Kovarianz-Matrix würde erneut geschätzt. Das könnte ggf. wiederholt werden.

Fellegi sieht die Möglichkeit einer Verknüpfung des Verfahrens mit der von Fellegi/Holt vorgeschlagenen Methode der Bestimmung der für Korrekturen ausreichenden kleinsten Mengen von Datenfeldern.¹⁾ Es könnte auch stichprobenhaft eine Kontrollerhebung durchgeführt werden, die größere Gewähr dafür bietet, wenigstens "optimale" Werte zu enthalten, die dann der Schätzung einer verzerrungsfreien Kovarianz-Matrix zugrunde gelegt werden könnten.

523 Eignung für die Volkszählung

Wenn Volkszählungen als Totalzählungen durchgeführt werden, weil Ergebnisse entweder in tiefer sachlicher Gliederung oder für kleine Regionen oder sogar kombiniert benötigt werden, verbieten sich die Verfahren der Parameterschätzung von selbst. Deren Zielsetzung, Gesamtheitstotale, Varianzen und Kovarianzen aus vorhandenen - für fehlerfrei erachteten - Informationen zu schätzen, weicht ab von der bei Volkszählungen, entweder den einzelnen Datensatz mit Hilfe seiner redundanten Informationen aus sich selbst heraus oder durch Zuhilfenahme anderer möglichst ähnlicher Datensätze zu bereinigen.

Hinzu kommt, daß die Skalierungseigenschaften der meisten Variablen in der Volkszählung eine "Rückrechnung" von den Totalen auf die Einzelwerte - die von den Autoren zumindest wahlweise vorgesehen ist - nicht gestattet. In der Volkszählung kommt es dagegen auf die Klassifikation der Einheiten in die durch die meist qualitativen Variablen gegebenen diskreten Kategorien an.

So kann der Vorschlag von Dalenius,²⁾ aus den Werten der Volkszählung ein System von Regressionsgleichungen zu entwickeln, in

1) Fellegi, I. P., Automatic Editing and Imputation of Quantitative Data, a.a.O., S.40.

2) Dalenius, T., Automatic Estimation of Missing Values in Censuses and Sample Surveys, in: Statistisk Tidskrift, Hrsg. Statistiska Centralbyrån, H. 7, Stockholm 1962, S.395-400.

dem fehlende Werte aus anderen Regressionsgleichungen von eigens zu diesem Zweck durchgeführten Stichprobenerhebungen zu schätzen wären, auch nur deskriptiv - im Hinblick auf die Bereinigung der Totalen - verstanden werden. Deskriptive Verfahren aber stellen keine Alternative zu den auf Gewinnung der Totalen durch Einzelfall-Aggregation abzielenden operativen Verfahren dar.

53 Einzelfall-Korrekturen

531 Stochastische Korrekturen

In diesem Abschnitt werden Korrekturen analysiert, die auf dem Zufallsprinzip basieren. Jedes Korrekturverfahren wird zuerst durch seine Wirkungsweise charakterisiert, es folgen Aussagen über die bisherige praktische Verwendung, schließlich wird eine Bewertung sowohl unter Genauigkeits- als auch unter technisch-organisatorischen Aspekten vorgenommen.

5311 Doppeln

Bei diesem Verfahren wird jeder Datensatz, in dem mindestens ein Wert durch das Kontrollsystem beanstandet ist, mit allen Werten zurückgewiesen und vollständig ersetzt durch einen zufällig ausgewählten, vom Kontrollsystem nicht reklamierten kompletten Datensatz. Das gesamte Datenmaterial gliedert sich also in Rückweisungen und Annahmen, wobei die Annahmen das Korrekturpotential darstellen.

Da das Verfahren in seiner normalen Ausgestaltung keine Rücksicht auf die Merkmalsstruktur der zu korrigierenden statistischen Einheit nimmt, kann es auch für Fälle eingesetzt werden, in denen die Informationsdichte ("Kräftefeld") eines Datensatzes seine Typisierung oder Klassifikation nicht zuläßt. Das ist z.B. gegeben, wenn die Auskunftsperson keine oder nur wenige Angaben gemacht hat, oder bei der Datenerfassung (z.B. Beleglesung oder Mikroverfilmung) keine oder nur wenige Informationen identifiziert werden konnten. Im Grenzfall genügen einige Ordnungsbegriffe (z.B. Regionalangaben) der Auskunftspersonen, um Pseudo-Datensätze auf den Magnetbändern anzulegen und somit die Voraussetzung für das Doppeln der Gesamtheit der Werte eines Datensatzes zu schaffen.

In USA waren 1960 0,4 % aller gezählten Personen gedoppelt worden. Ein Viertel davon (0,1 %) war auf nicht zu identifizierende FOSDIC-Belege¹⁾ zurückzuführen. Bei der Volkszählung 1970 in der

1) FOSDIC = Film Optical Sensing Device for Input to Computers. Bei dem Verfahren werden die durch geschwärzte Felder markierten Belege auf Mikrofilm übernommen. Das elektronische Gerät "liest" die Informationen und überträgt sie mittels Computer auf die Magnetbänder.

Bundesrepublik wurde im Anschluß an die automatischen Bestandskontrollen versucht, solche Totaleinsetzungen "strukturgerecht" (innerhalb desselben Haushalts, Gebäudes, Straße, Gemeinde) als manuelle Lösung durchzuführen. Dies sollte in Zukunft in das automatische Fehlerbereinigungsprogramm integriert werden.

Nach Meinung des Verfassers ist hier die Grenze der Rechtfertigung für "Korrekturen" erreicht. Der Schritt zur "Konstruktion" fiktiver Personen scheint nicht mehr weit. Wenn das Verfahren allerdings mit großer Verantwortlichkeit durchgeführt wird und keine Möglichkeit zu sinnvoller Typisierung oder Klassifikation der Datensätze besteht, dann sollte ein sehr kleiner Prozentsatz (weniger als 0,5 %) auf diese Weise "korrigiert" werden, weil trotz aller Bemühungen immer ein Rest nicht zu erfassender Fälle bleiben wird.

Zindler hat die Auswirkungen der Korrekturverfahren auf die Genauigkeit statistischer Ergebnisse unter Verwendung der Erkenntnisse der Stichprobentheorie berechnet.¹⁾ Das von ihm verwendete Modell hat folgende Prämissen:

- a) Die Bearbeitung der Fragebögen und die Korrektur der Felder sind zwei- bzw. dreistufige Zufallsauswahlen. In der ersten Stufe werden die Einheiten der Stichprobe aus der Gesamtheit zufällig ausgewählt. In der zweiten Stufe ist das Entstehen falscher Merkmalswerte dieser Einheiten vom Zufall abhängig. In der dritten Stufe unterliegen die zur Korrektur herangezogenen Ersatzwerte der Zufallsauswahl.
- b) Im Prozeß der automatischen Fehlerbereinigung werden die Datensätze zu Korrekturzwecken nach bestimmten Merkmalen abgefragt und Gruppen zugeordnet (Typisierung). In der Stichprobe befinden sich zufällig Einheiten dieser Gruppen. Vor der Korrektur werden die Datensätze gruppenweise sortiert.

1) Zindler, H. J., Die Auswirkungen maschineller Korrekturverfahren, in: Metrika, Internationale Zeitschrift für theoretische und angewandte Statistik, Vol. 11, Wien-Würzburg 1967, S.210-227. Der Beitrag war bereits vorher erschienen in: Szameitat, K., Zindler, H. J., The Reduction of Errors in Statistics by Automatic Corrections, in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.395-417.

- c) Alle Fehler entstehen im Prozeß der automatischen Fehlerbereinigung und werden durch die automatischen Kontrollen vollständig entdeckt.

Die Bewertung der Korrekturverfahren geschieht über die Berechnung des Anstiegs der Varianz infolge des Verfahrens. Z.B. ergibt sich bei einer Stichprobenauswahl von 5 % und Rückweisungen von Datensätzen in Höhe von 10 % eine Zunahme der Varianz von 21,1 %.

Zindler bezieht in seine Untersuchung die Auswirkungen der Gruppenbildung ein. Er kommt zu dem Schluß, daß - ähnlich wie bei geschichteten Stichproben - eine Gruppenbildung nicht für jedes Merkmal zu einer Verringerung des relativen Anstiegs der Varianz führt. Die Frage, nach welchen Merkmalen gruppiert werden soll und welche Anzahl von Gruppen optimal ist, müsse am Einzelfall entschieden werden.

Wenn die Fehler in den Datensätzen zufällig aufträten, könnte man das Ergebnis der Untersuchung als realistisch ansehen. Da bestimmte Gruppen von Berichtspflichtigen jedoch dazu neigen, bestimmte Fehler zu machen, andererseits bestimmte Merkmale fehleranfälliger als andere sind, bringt das Doppeln beträchtliche systematische Fehler mit sich.

Das Korrekturverfahren gilt als sehr grob, da bereits bei einem fehlerhaften Datenfeld alle übrigen Originalinformationen vernichtet werden. Es wird nur im äußersten Notfall verwendet. Aus technisch-organisatorischer Sicht ist vorteilhaft, daß keine Fehlerallokation notwendig ist und Iterationen, also Wiederholungen von Korrekturen, wegfallen.

Im Zusammenhang mit dem Doppeln ist noch das Verfahren "Doppeln und Streichen" zur Vereinfachung von Hochrechnungen zu erwähnen. Damit läßt sich eine Stichprobe, deren geplanter Auswahlatz durch Unter- bzw. Übererfassung nicht erreicht wurde, einem Aggregat anpassen. Beim Doppeln werden im Umfang der Untererfassung fehlende Einheiten zufällig ausgewählt und in die Aufbereitung einbezogen, beim Streichen entsprechend der Übererfassung Einheiten von der Aufbereitung ausgeschlossen. Das erspart die Bestimmung unter-

schiedlicher Hochrechnungsfaktoren, deren Anzahl bei tief gegliederten Tabellen erheblich sein kann. Meist sind die zu doppelnden oder zu streichenden Einheiten allerdings unter vielfältigen Kriterien klassifiziert, da die Hochrechnung an eine Reihe von Bedingungen geknüpft ist.

In der Bundesrepublik wurde dieses Verfahren bei der Volkszählung 1970 zur Anpassung des repräsentativen Teils der Erhebung verwendet.

5312 Cold Deck

Unter einem "Deck" wird eine auf Datenträger gespeicherte Matrix verstanden. Sie ist in ihrem Aufbau einer Tabelle vergleichbar, die der Darstellung statistischer Ergebnisse dient. Die Deckzellen enthalten jedoch nicht wie die Tabellenfelder absolute Häufigkeiten, sondern einzelne Merkmalsausprägungen, mit denen fehlerhafte Ausprägungen korrigiert werden.

Beim Cold Deck stammen diese Merkmalsausprägungen aus Häufigkeitsverteilungen vorhergehender Zählungen, aus vergleichbaren früheren Erhebungen oder Probeerhebungen zum aktuellen Projekt. Sie sollen für die jeweilige Deckzelle repräsentativ sein.

Als Beispiel soll ein "Altersdeck" dienen. Die Matrix soll als Vorspalte das Merkmal "Geschlecht" und im Kopf das Merkmal "Familienstand" aufweisen. Die Merkmalsausprägungen "männlich" und "weiblich" sowie "ledig", "verheiratet", "verwitwet" und "geschieden", miteinander kombiniert, ergeben 8 Deckzellen, in denen typische Alterswerte (z.B. arithmetische Mittel) gespeichert sind.

Wenn im Kontrollprozeß die o.g. Merkmalskombination zurückgewiesen wird und ein Fehler im Alter als für die Rückweisung verantwortlich angesehen wird, dann wird der Datensatz nach Geschlecht und Familienstand klassifiziert und der in der Deckzelle gespeicherte Alterswert als Ersatz übernommen. Aus der Literatur kann nicht eindeutig geschlossen werden, ob der

im Alter korrigierte Datensatz, wenn er weitere Fehler aufweist, anschließend anderen Decks zugeordnet wird, aus denen er weitere Ersatzwerte übernimmt, oder ob er nach der Zuordnung zu einer möglichst homogenen Gruppe von Datensätzen durch einen vollständigen "repräsentativen" Datensatz korrigiert wird. Letzteres ist wohl nur zu rechtfertigen, wenn eine Reihe von Klassifikationskriterien (nicht etwa nur "Geschlecht" und "Familienstand") eine "sachgerechte" Schichtung ermöglichen, damit die durch Übernahme des kompletten Datensatzes geänderten korrekten Originalinformationen den Typ des Datensatzes nicht verfälschen. Denkbar ist auch, aus dem "Mustersatz" in dieser Schicht zunächst einen Wert zu übernehmen; wenn bei der Iteration festgestellt wird, daß noch Inkonsistenzen im Datensatz vorhanden sind, einen zweiten Wert zu übernehmen. Im ungünstigsten Fall wäre die Korrektur erst nach Übernahme aller Werte abgeschlossen.

Die Qualität des Korrekturverfahrens hängt stark von der Schichtung des Materials ab. Der Aufwand an Maschinenzeit wird durch die Zahl der erforderlichen oder zugelassenen Iterationen bestimmt, die sich als nicht notwendig erweisen, wenn ein Datensatz vollständig durch einen unbeanstandeten ersetzt wird. Wenn keine Forschungsergebnisse darüber vorliegen, welches Merkmal in einer bestimmten Merkmalskombination vermutlich zur Rückweisung der Kombination geführt hat, dann bleibt oft nichts anderes übrig, als die Konzeption der Decks auf Vermutungen aufzubauen. Wenn aus der Merkmalskombination "Geschlecht", "Familienstand" und "Alter" nicht das Alter generell als am fehleranfälligsten gelten kann, dann besteht eigentlich keine Rechtfertigung für ein Altersdeck.

Nach Zindler ist es nicht schwierig, bei quantitativen Daten Mittelwerte als Ersatzwerte zu berechnen. Selbst bei Verzicht auf die Schichtung des Datenmaterials zeigten sich hier bereits gute Ergebnisse. Bei qualitativen Daten sei das jedoch unter Umständen nicht einmal bei extrem tiefer Schichtung zu erwarten. Wenn nicht für jedes Merkmal ein spezielles Deck eingerichtet würde, sei der systematische Fehler zu groß. Allerdings ist die Zuordnung eines fehlerhaften Datensatzes zu einer bestimmten Schicht unter Umständen schwer, wenn die Merkmale, die die Schicht bestimmen, selbst falsch sind oder fehlen. Im Rahmen der Beispiele deterministischer Korrekturen bei der Volkszählung 1970 wird der Versuch einer pragmatischen Lösung aufgezeigt.

Nun ein Beispiel für ein spezielles Deck zur Einsetzung des "Familienstandes". Angenommen, aus der Vergangenheit sei bekannt, daß 40 % der Männer verheiratet sind. Dann könnten in einem nur durch das Geschlecht charakterisierten Cold Deck in zufälliger Reihenfolge 10 Werte gespeichert sein, in 4 Fällen "verheiratet - ja", in 6 Fällen "verheiratet - nein". Immer wenn für einen Datensatz mit der Information "männlich" ein Familienstand gesucht wird (z.B. als Folge eines Blanks im Datenfeld "Familienstand" im Rahmen einer Einzelfeld-Kontrolle), dann wird das Deck hintereinander abgearbeitet.

Je intensiver die Schichtenbildung durchgeführt wird, umso mehr Speicherplatz wird verbraucht, da alle Decks im Speicher angelegt sein müssen. Es wäre z.B. daran zu denken, bei der Bestimmung der Decks zusätzlich regionale Unterschiede zu berücksichtigen. Das würde den Aufbau vieler Decks mit möglicherweise vielen Deckzellen bedingen. Oder es würden je Deckzelle "repräsentative" Werte von mehreren Personen gespeichert, die dann - systematisch oder zufällig ausgewählt - als Ersatzwerte herangezogen werden.

Alle Modifikationen haben das Ziel, die Nachteile dieses Verfahrens zu mildern, d.i. die künstliche Häufung um die "repräsentativen" Werte und somit die Beeinträchtigung der Streuungsverhältnisse. Ein weiterer Mangel ist gegeben, wenn die Korrekturwerte nicht den Verhältnissen der aktuellen Erhebung entsprechen.

Die Cold Deck-Methode wurde im US Census Bureau bereits bei der Volkszählung 1940 angewandt. In der Bundesrepublik wurden 1970 erstmals automatische Korrekturen durchgeführt; dabei kam das weiterentwickelte Hot Deck-Verfahren zum Einsatz.

5313 Hot Deck

Dieses Korrekturverfahren unterscheidet sich vom Cold Deck-Verfahren dadurch, daß die Ersatzwerte in den Deckzellen im Ablauf des Fehlerbereinigungsprozesses durch von den Kontrollen nicht beanstandete Werte oder Wertekombinationen ständig "erneuert" werden. So kann z.B. im o.g. Altersdeck in der Deckzelle "männlich"/"ledig" zufällig ein Alter von 50 Jahren gespeichert sein, das

als Ersatzwert in die beanstandete Wertekombination des nächsten Datensatzes eingesetzt wird.

Ähnlich wie beim Cold Deck-Verfahren ist der Literatur nicht zweifelsfrei zu entnehmen, ob sich innerhalb der Schichten, in die das Datenmaterial gegliedert ist, jeweils repräsentative Sätze ständig erneuern und als Korrekturpotential dienen oder Einzelwerte bzw. Wertekombinationen. Wenn sich Einzelwerte bzw. Wertekombinationen durch Substitution ständig verändern, wird das Verfahren als "modifiziertes" Hot Deck-Verfahren¹⁾ gekennzeichnet. Wenn nur einzelne Datenfelder eines Datensatzes zur Korrektur verwendet werden, kann es wieder zu Iterationen kommen. Außerdem müssen Kenntnisse über die Fehleranfälligkeit einzelner Merkmale in Kombinationen vorhanden sein.

Zu Beginn des Korrekturprozesses werden meist Cold Decks aufgebaut, da bereits der erste Datensatz des Datenbestands korrekturbedürftig sein kann. Um möglichst alle Deckzellen der Struktur des aktuellen Materials anzupassen, werden die Zellen durch einige tausend Werte "aufgeheizt".

Bei der Aufbereitung der 25 %-Stichprobe im Rahmen der Volkszählung 1960 in USA wurden z.T. sehr aufwendige Decks mit bis zu 1 500 Zellen gespeichert und im Ablauf jeweils erneuert.

Im Rahmen der stochastischen Korrekturen waren bei der Volkszählung 1970 in der Bundesrepublik 22 Decks entwickelt worden.²⁾ Es gelangten nur solche Merkmale als Korrekturpotential in die Decks, die weder bei der Zuordnung zu Schichten (Typen) noch bei der Zuordnung zu Decks beanstandet wurden.

Die von Fellegi/Holt vorgeschlagenen Korrekturen³⁾ basieren auch auf der Hot Deck-Methode. Grundidee ist, wie oben ausgeführt, daß je Datensatz die kleinste Zahl an Datenfeldern markiert ist, die zu ändern ist, damit der Datensatz keine Kontrollen mehr verletzt.

1) Vgl. Naus, J. I., Data Quality Control and Editing, a.a.O., S.113.

2) Übersicht über die Decks im Anhang S.176-179.

3) Fellegi, I. P., Holt, D., A Systematic Approach to Automatic Editing and Imputation, a.a.O., S.373-375.

Die Korrekturen heißen Sequential-Verfahren 1, Sequential-Verfahren 2 und Simultan-Verfahren. Beim Sequential-Verfahren 1 werden alle zu ändernden Datenfelder nacheinander behandelt. Es werden alle Kontrollen (Verknüpfung zulässiger Wertemengen) betrachtet, die zunächst das erste zu verändernde Datenfeld beinhalten sowie alle anderen nicht reklamierten Datenfelder des zu korrigierenden Datensatzes. Da die Werte in den nicht zu verändernden Datenfeldern des zu korrigierenden Datensatzes in den Wertemengen dieser Kontrollen enthalten sind, genügt es, wenn der erste Datensatz mit einem zulässigen Wert den unzulässigen ersetzt. In gleicher Weise werden alle zu verändernden Datenfelder sukzessive bereinigt.

Beim Sequential-Verfahren 2 wird im Unterschied zu Verfahren 1 die Forderung erhoben, daß die nicht zu verändernden Datenfelder des zu korrigierenden Datensatzes nicht nur innerhalb der Wertemengen der Kontrollen liegen, sondern daß der Datensatz, mit dem korrigiert wird, möglichst "ähnlich" ist. Die Autoren geben keinen Aufschluß darüber, wie sie diese "Ähnlichkeit" feststellen wollen. Sie sagen lediglich, daß man sich an den gültigen Wertemengen der Korrekturen orientieren müsse.

Beim Simultan-Verfahren müssen alle nicht zu verändernden Datenfelder des zu korrigierenden Datensatzes mit dem Datensatz, der zur Korrektur herangezogen wird, exakt übereinstimmen, ehe die zu verändernden Werte substituiert werden. Das kommt einer Klassifikation jedes einzelnen Datensatzes unter strengsten Bedingungen gleich.

Von Verfahren zu Verfahren wird die Struktur des zu korrigierenden Datensatzes stärker berücksichtigt, dabei natürlich der Suchaufwand größer. Die Autoren erkennen, daß aus Gründen der Wirtschaftlichkeit Verfahren 3 kaum durchzusetzen sein wird. Vermutlich wird auf Verfahren 1 zurückgegriffen werden müssen.

Die Hot Deck-Methode gilt als flexibel bei der Anpassung an die Struktur der Gesamtheit. Sie erhält die Verteilungseigenschaften einigermaßen gut und verwendet gleichzeitig aktuelle Daten. Im Vergleich zur Cold Deck-Methode wird dadurch Speicherplatz gespart, daß nicht alle Spezialitäten des Datenmaterials in Form besonderer Decks angelegt sein müssen. So ist z.B. der regionale

Schichtungseffekt im Datenmaterial meist bereits enthalten.

Zindler kommt zu dem Ergebnis, daß die Hot Deck-Methode etwa so wirkt wie eine Doppelung für jedes Merkmal einzeln. Wie weiter vorn ausgeführt wurde, bezieht sich das Doppeln jedoch auf vollständige Datensätze. Es sei zweifelhaft, ob dieselbe Schichtenbildung für alle Merkmale vorteilhaft sei, auch die Frage nach geeigneten Decks müßte das Ergebnis empirischer Untersuchungen von Fall zu Fall sein. Unter den strengen Modellannahmen hinsichtlich der Zufallseigenschaften der Fehlerentstehung weist er nach, daß die Ergebnisse bereits ohne Gruppenbildung günstig seien.

Eigene Einblicke in den Korrekturablauf der Volkszählung 1970 - obwohl aus Kostengründen nicht quantifiziert - zeigen,¹⁾ daß in beträchtlichem Maße manuelle "Überkorrekturen" notwendig waren, um die schlimmsten Auswirkungen von Serieneinsetzungen, die sich ergaben, wenn mehrere Datensätze hintereinander durch denselben im Deck befindlichen Wert oder Wertekombination ersetzt werden mußten, zu beseitigen. Verbesserungen könnten hier erzielt werden, wenn je Deckzelle mehrere Werte gespeichert wären, auf die im Bedarfsfall zufällig oder systematisch zugegriffen werden könnte.²⁾ Natürlich stehen bei der Konzeption der Decks deren Anzahl und Umfang in Konkurrenz zu Wirtschaftlichkeitsüberlegungen.

5314 Monte Carlo-Methode

Unter der Monte Carlo-Methode in ihrer allgemeinen Bedeutung wird die Lösung analytischer Probleme durch Zufallsexperimente verstanden. Die Probleme werden in Form stochastischer Modelle simuliert. Die Ergebnisse werden durch statistische Auswertung

1) Vgl. S.130.

2) Bericht über das Expertentreffen zur Beratung von Volkszählungsfragen, Hrsg. Österreichisches Statistisches Zentralamt, Wien 1968, S.8.

der Experimente und Schätzung der Parameter erzielt.¹⁾

Als automatisches Korrekturverfahren verstanden, beruht diese Methode auf der Zufallsauswahl von Ersatzwerten aus den Häufigkeitsverteilungen der beobachteten Variablen. Die Verteilungen werden aus unbeanstandeten Werten vorhergehender Zählungen oder Probeerhebungen zu aktuellen Zählungen entwickelt. Es wird unterstellt, daß die für die Korrektur benutzten Werte "wahre" Werte sind. Mit Hilfe eines Zufallszahlengenerators werden Zufallszahlen erzeugt, die auf die empirische Häufigkeitsverteilung angewandt werden und dort zur Bestimmung des Wertes einer Variablen als in den fehlerhaften Datensatz zu übernehmenden Ersatzwert führen.

Als Anwendungsbeispiel soll nochmals die bereits oben erwähnte Alterskorrektur herangezogen werden. Es handelte sich um eine Merkmalskombination aus "Geschlecht", "Familienstand" und "Alter". Für jede der 8 Merkmalskombinationen aus "Geschlecht" und "Familienstand" ist eine eigene Altersverteilung zu entwickeln und zu speichern. Ein Datensatz mit fehlenden oder beanstandeten Alterswerten wird zunächst auf Grund der Abfrage nach "Geschlecht" und "Familienstand" der relevanten Verteilung zugeordnet, dann wird in Abhängigkeit von der Zufallszahl ein bestimmtes Alter "gezogen" und als Ersatzwert in den Datensatz übernommen.

Die Monte Carlo-Methode überwindet Beschränkungen als Folge der Bestimmung von Abschneidegrenzen, d.h. sie bewahrt die Verteilungseigenschaften. Dabei wird natürlich angenommen, daß die zurückgewiesenen Werte entsprechend den unbeanstandeten verteilt sind. Das ist in manchen Fällen zu bezweifeln. Zindler berechnet, daß der relative Varianzanstieg als Folge dieses Korrekturverfahrens selbst bei Verzicht auf die Schichtung des Materials relativ klein ist. Allerdings wird vorausgesetzt, daß der Mittelwert der Merkmalswerte in der Grundgesamtheit nicht allzusehr

1) Vgl. Koelle, H. H., Die Anwendung der Simulation als Entscheidungshilfe, in: IBM Nachrichten, 21. Jg., H. 208, Verlag IBM Deutschland, Sindelfingen 1971, S.873-879; Bauknecht, K., Nef, W. (Hrsg.), Digitale Simulation, in: Lecture Notes in Operations Research and Mathematical Systems, H. 51, Springer, Berlin-Heidelberg-New York 1971, S.10-34; Kohlas, J., Monte Carlo Simulation in Operations Research, in: Lecture Notes in Economics and Mathematical Systems, H. 63, Springer, Berlin-Heidelberg-New York 1972.

abweicht vom Mittelwert der Ersatzwerte, und daß die Varianz der Ersatzwerte höchstens gleich ist der der Merkmalswerte. Damit die Gruppenbildung bessere Ergebnisse bringt, muß der "Schichtungseffekt" relativ groß sein.

Wenn die Datensätze in zufälliger Reihenfolge korrigiert werden, wird der Erfolg der Monte Carlo-Methode als ebenso groß erachtet wie der des Hot Deck-Verfahrens. Bei "versteckter" Anordnung der Datensätze - wie das üblicherweise der Fall ist - gilt das Hot Deck-Verfahren jedoch als überlegen.

Aus technisch-organisatorischer Sicht ist das Hot Deck-Verfahren besser in das Fehlerbereinigungssystem zu integrieren, da Kontrollen und Korrekturen in einem Maschinenlauf durchgeführt werden können. Bei der Monte Carlo-Lösung werden entweder historische Verteilungen benötigt, oder es sind getrennte Maschinenläufe für Kontrollen und Korrekturen erforderlich. Nur in einem Spezialfall kann unterstellt werden, daß die zu kontrollierenden Daten einer bestimmten Verteilung folgen: die Endziffern bei Geburtsjahren einer Rechteckverteilung.¹⁾ Hier können die Korrekturen parallel zu den Kontrollen ablaufen.

Es erscheint sehr fragwürdig, aktuelles Datenmaterial an historische Werte oder Verteilungen anzupassen.²⁾ Unter Genauigkeitsaspekten ist die Entwicklung der Verteilungen aus aktuellem Material in einem Maschinenlauf und die Korrektur aus diesen Verteilungen in einem zweiten Lauf besser. In der Praxis sind beide Lösungen schwer zu verwirklichen: die Integration historischer Werte in die Fehlerbereinigung deshalb, weil das Frageprogramm von Zählung zu Zählung variiert; die Verarbeitung in zwei Phasen, weil die Aktualität der Ergebnisse gefährdet ist. Diese Probleme tauchen bei deterministischen Korrekturen nicht auf. Im folgenden werden zuerst zwei für quantitative Merkmale geeignete Verfahren vorgestellt, ehe die für Volkszählungen typischen behandelt werden.

-
- 1) Vgl. Nordbotten, S., Automatic Editing of Individual Statistical Observations, a.a.O., S.30.
 - 2) Pritzker, L., Ogus, J., Hansen, M. H., Computer Editing Methods. Some Applications and Results, in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.446.

532 Deterministische Korrekturen5321 FUNOM (Full Normal Modification)

Tukey hat FUNOM in Zusammenhang mit der FUNOR-Kontrolle entwickelt.¹⁾ Die Fehlerbeseitigung sieht einmal vor, die durch FUNOR zurückgewiesenen Ausreißer an den Median anzupassen. Zum anderen werden weitere Ausreißer - selbst wenn sie nicht beanstandet wurden - "zensoriert" oder "winsorisiert". Zensurierung heißt, in einer geordneten Stichprobe eine vorbestimmte Anzahl größter und kleinster Werte zu eliminieren. Winsorisation bedeutet, diese vorbestimmte Anzahl größter oder kleinster Werte durch die jeweils nächstgelegenen zu ersetzen. In beiden Fällen werden die Parameter bei unverändert gebliebenem Stichprobenumfang geschätzt. Durch Variation der Vorgabewerte A und B bei FUNOR²⁾ werden nicht nur die Annahme- bzw. Rückweisungsgrenzen der Kontrolle bestimmt, sondern auch die Abschneidegrenzen durch die FUNOM-Korrektur.

Scott hat die Technik der Winsorisation näher untersucht und kommt zu dem Schluß, daß die Anpassung von 3 - 6 % der beobachteten Werte - allerdings symmetrisch - eine Verteilung nicht ernstlich beeinträchtigt.

5322 Funktionale Korrekturen

Die Funktionallösung bei Korrekturen ist das Gegenstück zur Funktionallösung bei Kontrollen.³⁾ Funktionale Abhängigkeiten zwischen den Variablen können sowohl zu Kontrollzwecken als auch zu Korrekturzwecken genutzt werden. So kann z.B. bei der Additionskontrolle, wenn die Summe der Komponenten mit dem Totalwert nicht übereinstimmt, nach dem vergeblichen Versuch, durch Änderung einzelner Komponenten den Totalwert zu erreichen, unter bestimmten Voraussetzungen als Totalwert die rechnerische Summierung der Komponenten übernommen werden. Als Konsequenz der Quotienten-Kontrolle im Rahmen der Fehlerbereinigung bei der

1) Vgl. Naus, J. I., Data Quality Control and Editing, a.a.O., S.115.

2) Vgl. S.26/27.

3) Vgl. Nordbotten, S., Automatic Editing of Individual Statistical Observations, a.a.O., S.30-32.

Außenhandelsstatistik ist eine Funktionallösung denkbar, die z.B. den Mittelwert für x_1 (Wert) und den Mittelwert für x_2 (Menge) speziell für jede Warengruppe als Ersatzwerte einsetzt. Das setzt Vorüberlegungen zur Konfidenz entweder des Merkmals "Wert" oder des Merkmals "Menge" voraus. Entweder kann historisches Wissen zur Fehlerallokation herangezogen werden, oder das Merkmal mit der größten Abweichung vom Mittelwert (bezogen auf die jeweilige Warengruppe) aus dem aktuellen Datenmaterial kann als Ursache für die Rückweisung des Quotienten angesehen werden.

Die folgenden Funktionallösungen können als typisch gelten:

$$f_1(x_1'', x_2'') = x_1'' - \frac{\bar{x}_1}{\bar{x}_2} \cdot x_2'' = 0 \quad \begin{array}{l} x_i'': \text{bereinigte Werte} \\ x_i': \text{beobachtete Werte} \end{array}$$

$$f_2(x_1'', x_2'') = x_1'' + x_2'' - \bar{x}_1 - \bar{x}_2 = 0 \quad x_i: \text{"wahre" unbekannte Werte}$$

Wenn x_1 (Wert) als falsch erkannt ist, wird als Ersatzwert

$$x_1'' = \frac{\bar{x}_1}{\bar{x}_2} \cdot x_2' \text{ eingesetzt}$$

Ein zurückgewiesener "Wert" wird also mit Hilfe des Quotienten aus einem "guten" Durchschnitts- "Wert" und einer "guten" Durchschnitts- "Menge" korrigiert.

Für den Fall, daß sowohl "Wert" als auch "Menge" beanstandet werden, dienen beide Gleichungen zur Bestimmung von Ersatzwerten.

Gegen diese Art von Korrektur ist natürlich einzuwenden, daß sie die Variabilitätsstruktur der statistischen Gesamtheit u.U. beträchtlich stören kann. Stünden mehrere Schätzwerte als Ersatzwerte zur Verfügung, könnte ein stochastisches Korrekturverfahren eingesetzt werden.

Ersatzwerte für fehlende oder beanstandete Werte können auch aus Regressionsgleichungen gewonnen werden. Die Schätzungen können entweder auf den nicht beanstandeten kompletten Datensätzen beruhen oder auf der Gesamtheit besetzter Datenfelder, gleichgültig, ob sie aus vollständigen oder unvollständigen Datensätzen stammen.

Die Tatsache, daß durch die Schätzungen primär Datensätze komplettiert werden und erst in zweiter Linie die Kovarianzmatrix - quasi als Qualitätsmaßstab - berechnet wird, läßt die Zuordnung dieser Korrekturart zu den Einzelfall-Korrekturen als gerechtfertigt erscheinen.

In einer Studie von Buck¹⁾ werden nur die unbeanstandeten vollständigen Datensätze zur Regressionsschätzung herangezogen. Untersucht werden quantitative Merkmale an Zuckerrüben. Es wurden 72 Stichproben mit je 4 Meßwerten gezogen. Aus den insgesamt 288 Meßwerten wurden 35 nach einem Zufallsverfahren unterdrückt, so daß in 29 Stichproben 1, 2 oder 3 Meßwerte fehlten. Mit Hilfe der multiplen Regressionsrechnung wurden dann die fehlenden Werte aus den kompletten Datensätzen geschätzt.

Die Kovarianzmatrix weist nach Vornahme der Korrekturen der auf diese Weise gewonnenen Ersatzwerte deutlich ungünstigere Werte auf als im Falle der ursprünglichen Matrix ohne Veränderung durch Korrekturen. Sie ist allerdings nur geringfügig schlechter als für den Fall, in dem alle besetzten Datenfelder für die Regressions-schätzung zugrunde gelegt wurden. Buck empfiehlt ähnliche Studien an höher korreliertem Material. Ein weiterer Lösungsansatz von Freund/Hartley²⁾ stellt ein vollständiges Fehlerbereinigungsprogramm, bestehend aus mehreren Kontroll- und Korrekturschritten dar. Das Verfahren bezieht sich ebenfalls auf quantitative Variable metrischer Skalierung. Der Ablauf vollzieht sich in drei Schritten.

- a) Zunächst erfolgt eine relativ grobe Einzelfeld-Kontrolle, die auch wenig differenzierte Einsetzungen in die beanstandeten Datenfelder nach sich zieht.
- b) Dann schließt sich eine Kontrolle von Wertekombinationen mit relativ einfacher Fehlerbeseitigung an. Die Art der Fehlerbeseitigung wird nicht näher spezifiziert. Historische Daten, "ähnliche" Beobachtungen aus der laufenden Zählung und für konsistent gehaltene Ersatzwerte sollen Verwendung finden.

1) Buck, S. F., A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with Electronic Computer, a.a.O., S.302-306.

2) Freund, R. J., Hartley, H. O., A Procedure for Automatic Data Editing, in: Journal of the American Statistical Association, Vol. 62, No. 319, Menasha, Wisc. 1967, S.341-352.

- c) Falls noch Kontrollverletzungen auftreten - das dürfte nach Meinung des Verfassers die Regel sein, da die Identitätsbedingungen bislang nicht geprüft wurden - wird ein Korrekturverfahren nach der "Methode der kleinsten Quadrate" vorgeschlagen.

Ziel ist es, eine Reihe linearer "Konsistenzbedingungen" (das sind Identitätsgleichungen) durch Änderung einiger Datenfelder - zumindest näherungsweise - zu erfüllen. Das Konzept geht von der Annahme aus, daß einige Gleichungen bedeutsamer sind als andere. Diese müssen mit größerer Genauigkeit erfüllt werden. Gleichzeitig gilt, daß einige Datenfelder als verlässlicher angesehen werden als andere und folglich in geringerem Ausmaß zu ändern sind. Die "korrigierten" Werte sollen möglichst "in der Nähe" der ursprünglichen liegen. Das soll erreicht werden durch Minimierung der gewogenen Summen der quadrierten Differenzen zwischen ursprünglichen und "korrigierten" Daten plus der gewogenen Summen der Quadrate der Diskrepanzen zwischen den Konsistenzgleichungen. Die Gewichte sind Ausdruck der Bedeutung der Bedingungen und/oder der Zuverlässigkeit der Daten.

$$\sum_j^n w_j \cdot (x_j - y_j)^2 + \sum_i^m u_i \cdot \left(\sum_j^n a_{ij} x_j \right)^2$$

x_j : j-te korrigierte Angabe

y_j : j-te ursprüngliche Angabe

w_j : Gewicht der j-ten Angabe; großes Gewicht weist auf größere Zuverlässigkeit der Angabe hin

$\sum_j^n a_{ij} x_j$: i-te Konsistenzgleichung der Form $\sum_j^n a_{ij} x_j = 0$

u_i : Gewicht der i-ten Konsistenz; großes Gewicht bedeutet wichtige Konsistenz

Die auf diese Weise bestimmten Ersatzwerte haben die Eigenschaft, in relativer Nähe zu den Ursprungswerten zu liegen und die Konsistenzgleichungen relativ gut zu erfüllen.

Das Problem liegt zweifellos in der Bestimmung der Gewichte. Im Hinblick auf die Variablen handelt es sich um nichts anderes als um das Problem der Fehlerallokation. Die Autoren meinen, daß die Gewichte Ausdruck der Variabilität der Daten der "gewöhnlichen" Zuverlässigkeit beim Berichten und der Zuverlässigkeit in bestimmten Fragebögen bzw. Datensätze sein sollten.

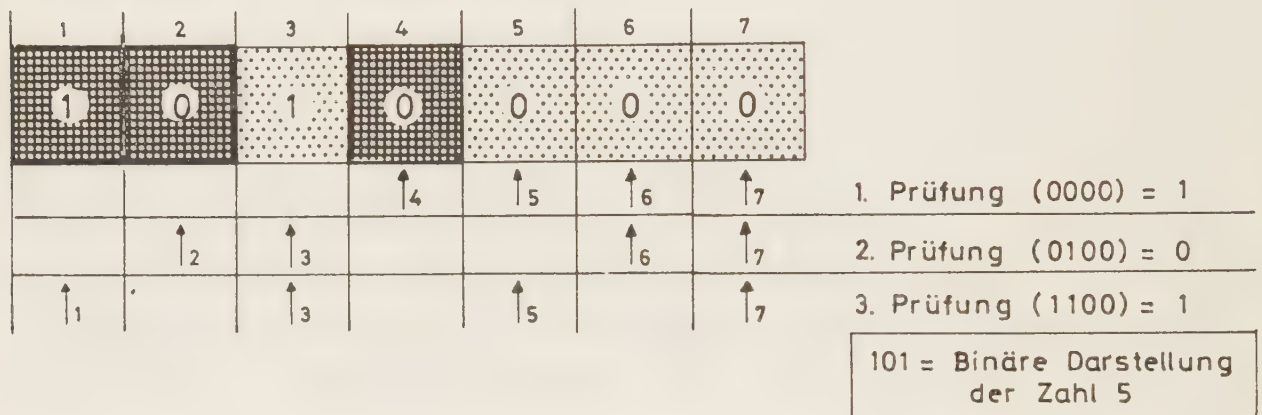
Die Gewichte für die Konsistenzgleichungen sollten die Reihenfolge zum Ausdruck bringen, in der diese Gleichungen mit abnehmender Anforderung an die Genauigkeit zu erfüllen sind.

Das Zahlenbeispiel der Autoren stammt aus der Landwirtschaftszählung 1961 in USA. Es wurden acht Varianten von Korrekturen durchgeführt. Maßstab für die Wirksamkeit jeder Korrekturart war die Differenz zwischen den fehlerfreien und "korrigierten" Angaben sowie die Standardabweichungen der "korrigierten" Angaben. Bei den von den Autoren gewählten Gewichten brachten bereits die Korrekturen im Anschluß an die Einzelfeld-Kontrolle gute Ergebnisse. Allerdings waren die Konsistenzbedingungen nach diesen Einsetzungen nur unzureichend erfüllt. Die Kleinst-Quadrate-Methode verbesserte die Daten noch beträchtlich.

Wenn es gelingen sollte, die Bestimmung der Gewichte in befriedigender Weise vorzunehmen, kann diese Methode der Fehlerbereinigung für Identitätsgleichungen flexibel eingesetzt werden. Aus technisch-organisatorischer Sicht erscheint der Programmieraufwand nicht hoch zu sein. Für Volkszählungen mit überwiegend qualitativen Merkmalen ist dieses Verfahren ungeeignet.

5323 Korrektur von Signierbegriffen53231 Einzelfehler-Korrekturcode

Deterministische Korrekturen können - ähnlich dem Kontrollziffern-Verfahren - auch mittels redundanter Ziffern durchgeführt werden. Das Prinzip soll am Beispiel des Einzelfehler-Korrekturcodes erläutert werden.

DER EINZELFEHLER - KORREKTURCODE

Die Bitpositionen 3, 5, 6 und 7 dienen der Verschlüsselung der Information, die übrigen der Odd-parity-Kontrolle. Es laufen drei Prüfungen ab; die erste für die Bitpositionen 4, 5, 6 und 7, die zweite für 2, 3, 6 und 7, die dritte für 1, 3, 5 und 7. Wenn die Anzahl der Einsen in einer Gruppe von Bitpositionen ungerade ist, wird 0 vergeben, wenn sie gerade ist dagegen 1. Als Ergebnis der drei Prüfungen in der Abbildung ergibt sich die Prüfziffer 101. Das ist die binäre Darstellung der Dezimalzahl 5, die angibt, daß die 5. Bitposition falsch übertragen wurde. Die Information muß 1100 lauten anstatt 1000. Die Korrektur wird durchgeführt.

Bose hat für ein Oktal-Zahlensystem Korrekturen für die beiden häufigsten Fehlerarten - eine Ziffer des Codes ist falsch bzw. zwei aufeinander folgende Ziffern sind vertauscht - entwickelt. Das setzt einen siebenstelligen Signierbegriff voraus, an den drei

- 1) Brennan, J. F., Computer "reparieren" sich selbst - Entwicklung von Verfahren zur automatischen Fehlersuche und -korrektur, a.a.O., S.8 und 10.

Ziffern als Korrekturbegriff angehängt werden. Wie oben bereits ausgeführt,¹⁾ gibt es bei den Signierbegriffen der Volkszählung keine Kontrollziffern.

53232 Beispiele für deterministische Lösungen aus Volkszählungen

Die Beispiele betreffen Korrekturen der Merkmale "Wirtschaftszweig" und "Beruf" durch das US Census Bureau sowie zwei Unterprogramme von zentraler Bedeutung im Rahmen der automatischen Fehlerbereinigung der Volkszählung 1970 in der Bundesrepublik.

Korrektur der Merkmale "Wirtschaftszweig" und "Beruf"

Die Schlüsselverzeichnisse für beide Merkmale enthielten Signierbegriffe, die aus jeweils drei Ziffern bestanden. Die erste Ziffer bestimmte die Ordnung, die zweite Ziffer die Gruppe, die dritte Ziffer die Klasse. Wenn ein Signierbegriff abgefragt wurde, dessen erste Ziffer in dem betreffenden Datenfeld fehlte, wurde dieser den Fällen "ohne Angabe" zugeordnet. Die Information wurde später mit Hilfe stochastischer Verfahren ergänzt. Eine fehlende zweite Ziffer wurde in der Weise ergänzt, daß die Ziffer der Gruppe übernommen wurde, die sich innerhalb der gegebenen Ordnung als die in der Vergangenheit am stärksten besetzte erwiesen hatte. Wenn die dritte Ziffer fehlte, wurde der Fall in die größte Klasse innerhalb der gegebenen Gruppe gesteuert. Bei ungültigen Codes wurde versucht, durch Untersuchung eines Zifferntauschs eine Zuordnung herzustellen. Vollständig fehlende Schlüssel wurden auf Grund kombinierter Abfragen der vorhandenen demographischen und ökonomischen Merkmale ergänzt. Dabei ist "mit hoher Wahrscheinlichkeit die richtige Ordnung eingesetzt worden. Da stärker abgrenzende Merkmale aus der Informationsstruktur aber nicht zu gewinnen waren, konnte für die Gruppe oder gar Klasse keine Garantie übernommen werden"²⁾.

1) Vgl. S.46.

2) 1970 Census of Population, Job Content Computer Edits and Allocation Specifications; Erläuterung des Programms in einem Schreiben von H. P. Miller, Leiter der Statistical Methods Division des US Bureau of the Census, an den Verfasser (Washington, 20. Dezember 1971).

Korrektur von Mehrfach-Antworten

Im allgemeinen ist auf eine Frage des Fragebogens nur eine Antwort zulässig. Mehrere Antworten führen zu inkonsistenten Informationen. Fragestellungen, in denen die Begriffe "überwiegend" und "hauptsächlich" vorkommen, gelten als besonders fehleranfällig, weil die Auskunftspersonen hier nicht selten Entscheidungen ausweichen. Nach einer Auszählung von 59 040 Fragebögen in einer Probezählung zur Volkszählung 1970 hatten sich bei der Frage nach dem hauptsächlich benutzten Verkehrsmittel für den Weg zur Arbeit oder Schule in 2,6 % aller Fälle (Auswertung erfolgte nur für Erwerbstätige) Mehrfach-Antworten ergeben.¹⁾ Immerhin wurden auch 0,6 % Fälle registriert, die bei der Frage nach dem Geschlecht sowohl "männlich" als auch "weiblich" als Antwort markiert hatten.

Die Erfahrung mit früheren Zählungen hatte gezeigt, daß Mehrfach-Antworten in der überwiegenden Zahl der Fälle Doppelantworten waren. Daher waren bei drei und mehr Antworten zu einer Frage alle Informationen gelöscht und Einsetzungen der sich anschließenden Fehlerbereinigung überlassen worden.

Aus den Regeln für die automatische Behandlung der Mehrfach-Antworten sei eine beispielhaft erläutert.²⁾ Im Falle des "hauptsächlich benutzten Verkehrsmittels" sah das Konzept die Vergabe von Prioritäten nach der Reichweite des Verkehrsmittels vor. Das Auto rangierte an der Spitze. Dem lag die Prämisse zugrunde, daß neben dem Auto selten ein anderes Verkehrsmittel benutzt wird. Im Einzugsbereich von Großstädten, bei einem relativ gut ausgebauten Netz öffentlicher Verkehrsmittel und relativ niedrigen Tarifen (wie etwa im Großraum Hannover) trifft das allerdings nicht zu. Dort wird das Auto oft nur für den Weg zwischen Wohnung und nächstgelegenen Bahnhof benutzt, so daß der Eisenbahn Priorität zukäme. Die regional nicht differenzierte Regelung, U-Bahn und Hochbahn an den Schluß der Prioritätsskala zu setzen, war in den Städten wie Berlin und Hamburg vermutlich fehleranfällig; gerade hier

- 1) Schreiben des Statistischen Bundesamtes, Wiesbaden, vom 10.10.196 VIII V 2/00, Br/Gr: Manuelle und maschinelle Auswertung der 6. Probeerhebung zur Volkszählung 1970, Anlage 3, Tabelle 5.
- 2) Schreiben des Statistischen Bundesamtes, Wiesbaden, vom 19.2.1970 VIII V 4/00, Schk/Br: Verfahren bei Mehrfach-Markierungen.

konnte mit einer Häufung von Mehrfach-Antworten gerechnet werden.

Nach Meinung des Verfassers hätte die Behandlung von Mehrfach-Antworten in der Phase der Fehlerbereinigung in die Kombinationskontrollen integriert werden müssen, da die Zulässigkeit einer Merkmalsausprägung in den meisten Fällen nur durch Betrachtung von Merkmalskombinationen entschieden werden kann. Es ist allerdings zu bedenken, daß das Kontrollprogramm für diese Fälle nur eine schmale logische Basis zur Einschränkung sinnvoller Kombinationen hat. Das hier praktizierte Verfahren, bereits im Rahmen der Datenerfassung deterministisch zu korrigieren, ist unter technisch-organisatorischen Aspekten sicher vorteilhaft, weil es Maschinenzeit spart. Es stellt unter qualitativen Gesichtspunkten jedoch keine befriedigende Lösung dar.

Schichtung der Datensätze

Die Schichtung wurde mittels Zuordnung jedes Datensatzes zu einem "Typ" vorgenommen. Auf dieser Zuordnung basierte in der Korrekturphase die Übernahme von Werten und Wertekombinationen aus Decks, die nach gleichen Kriterien gestaltet waren.

Aus technisch-organisatorischer Sicht hat die Typenbildung den Vorteil, daß mit der maschinellen Abfrage eines Typs eine Reihe von Merkmalskombinationen gleichzeitig verarbeitet werden kann. Typen wurden ausschließlich zum Zweck der Fehlerbereinigung konstruiert.

Bei der Volkszählung 1970 in der Bundesrepublik wurde die Bevölkerung in 5 Schichten gegliedert:

- Typ A: Kleinkind
- Typ B: Schulkind
- Typ C: Schüler, Studierender
- Typ D: Erwerbstätiger, Soldat
- Typ E: Nichterwerbsperson, Erwerbsloser

Die wichtigsten typendefinierenden Merkmale waren "Alter", "Schul-

besuch", "Schulabschluß" und "Stellung im Erwerbsleben". Um die Zuordnung mit größerer Verlässlichkeit vornehmen zu können, wurden weitere Merkmale zur Charakterisierung des Typs herangezogen: "Familienstand", "Stellung innerhalb des Haushalts", "überwiegender Lebensunterhalt" und ergänzende Angaben zur Erwerbstätigkeit.

Die Typenzuordnung erfolgte in Abhängigkeit von der Informationsdichte der Datensätze auf drei verschiedene Arten: fehlerfreie Zuordnung, Fehlertabelle, Zwangszuordnung.

Beispiel: Definition des Typs B

Alter: Januar bis Mai 1965, 1964 - 1956,
Juni bis Dezember 1955

Familienstand: ledig

Stellung innerhalb
des Haushalts: (Schwieger-)Sohn, Tochter des Haushaltsvorstands (HV); Enkel, Urenkel des HV; andere mit dem HV verwandte oder verschwägte Person; mit dem HV nicht verwandte Person

Überwiegender
Lebensunterhalt: eigene Rente, Pension; Unterhalt durch Eltern, Ehemann usw.; eigenes Vermögen, Vermietung, Zinsen, Altenteil; Sozialhilfe und sonstige Unterstützungen

Schulbesuch: Volksschule; Realschule; Gynmasium; Berufsfach-/ Fachschule

Schulabschluß: leer

Stellung zum
Erwerbsleben: nicht erwerbstätig

Verkehrsmittel/
Zeitaufwand: eine Angabe erforderlich

Weitere Angaben zur
Erwerbstätigkeit: leer

Ein Datensatz, der dieses Zuordnungsschema fehlerfrei durchlief, galt als einwandfrei zugeordnet. Wenn die Informationsdichte der Datensätze abnahm, wurde die Zuordnung der Person über eine Fehlertabelle versucht. Dem lag die Hypothese zugrunde, daß bis zu drei der typenrelevanten Merkmale als Ersatzwerte aus der Informationsstruktur des Datensatzes abgeleitet oder konsistent in diesen eingefügt werden könnten.

Eine Analyse der Fehlertabelle zeigt, daß sie als Konstruktionsprinzip das nominale Über"gewicht" von z.B. vier als richtig erkannten Merkmalen gegenüber drei als falsch erkannten enthielt. Selbst das Alter, das bei der Feldarbeit zur Volkszählung 1970 in den gemeindlichen Nachkontrollen und im Signierprozeß Priorität vor allen anderen Merkmalen hatte, büßte diese bei der Typenzuordnung ein. Es konnte zu Zuordnungen kommen, die im Gegensatz zum Alter standen. Prinzipien der Fehlerallokation sind hier noch nicht berücksichtigt.

Als letzter Schritt, wenn die beiden vorhergehenden erfolglos durchlaufen waren, folgte die Zwangszuordnung. Hier genügten bereits einzelne Abfragen, etwa nach dem Alter, dem Schulbesuch bzw. Schulabschluß oder der Erwerbstätigkeit, zur Aufnahme in die jeweilige Schicht. Das System war so konzipiert, daß jeder Datensatz spätestens mit Hilfe der "Zwangszuordnung" typisiert war.

Die Schwierigkeiten, die sich dadurch ergeben, daß schichtendefinierende Merkmale fehlen oder falsch sein können, sollten ursprünglich bei der Volkszählung in ähnlicher Weise überwunden werden wie bei der Gebäude- und Wohnungszählung 1968: durch Bewertung qualitativer Merkmale.¹⁾ So sollten die Auskunftspersonen

1) Vgl. Deininger, R., Kerler, E., Maschinelle Datenprüfung in der amtlichen Statistik - Automatische Korrekturverfahren bei der Gebäude- und Wohnungszählung 1968, S.299.

bei der Volkszählung mit ihren vorhandenen Merkmalen an "Idealtypen" gemessen werden. Die Informationsstruktur jedes Datensatzes sollte unter Berücksichtigung der vorhandenen Merkmale mit Konsens- bzw. Dissenswerten zu diesen Idealtypen ausgestattet werden. Die Zuordnung sollte entweder bei höchstem Konsens oder bei niedrigstem Dissens herbeigeführt werden. Das Bewertungsschema hatte sechs Typen vorgesehen. Die typendefinierenden Merkmale waren im wesentlichen die oben genannten. Nach dem Beitrag, den eine vorhandene Merkmalsausprägung zur Bestimmung des Typs lieferte bzw. den Auswirkungen, die bei nicht vorhandenen Angaben unterstellt wurden, waren 0, 2, 5, 10 oder 15 Strafpunkte vergeben worden. Nach diesem Schema erfolgte eine maschinelle Zuordnung von 2 617 Fällen. Zum Vergleich wurde von Fachleuten eine manuelle Zuordnung mit folgendem Ergebnis durchgeführt:

Typ	Manuelle	Maschinelle
	Zuordnung	
A	313	309
B	366	331
C	55	50
D	1066	999
E	5	5
F	812	698
Fehlzuordnung	-	225
Insgesamt	2617	2617

Die Abweichungen, besonders bei den Typen B und F wurden für so schwerwiegend gehalten, daß davon abgesehen wurde, diesen Weg weiter zu verfolgen.

Die Probleme optimaler Schichtung des Datenmaterials für Zwecke automatischer Fehlerbereinigung sind noch ungelöst. Hier könnte die Analyse von Kontingenztafeln mit Hilfe des χ^2 -Unabhängigkeitstests oder χ^2 -Homogenitätstests möglicherweise helfen bei der Klassifikation der statistischen Einheiten in Gruppen bzw. Deckzellen. Es könnten Maße für die Homogenität der Einheiten innerhalb einer Gruppe bzw. Deckzelle gefunden oder getestet

werden, ob die Unterschiede zu anderen Gruppen signifikant sind. Da bei jeder statistischen Erhebung der Fall gegeben sein dürfte, daß schichtdefinierende Merkmale fehlen oder falsch sind, wäre die oben ausgeführte Lösung von Fellegi/Holt, ein Minimum an Datenfeldern mit nicht beanstandeten Werten so zu ergänzen, daß alle Kontrollen unverletzt bleiben, intensiver praktischer Erprobung zu unterziehen.

6 Analyse des Systems der Fehlerbereinigung bei der Volkszählung 1970 in der Bundesrepublik

Im Rahmen der Fehleridentifikation und Fehlerbeseitigung sind bereits exemplarisch einige Teile des Gesamtprogramms behandelt worden. In diesem Teil der Arbeit ist beabsichtigt, den sachlichen Zusammenhang der wichtigsten Kontrollen und Korrekturen, der zugleich die zeitliche Verzahnung der Verfahren zeigt, auszuführen und über eigene Erfahrungen zur Wirksamkeit der Fehlerbereinigung zu informieren. Der Umfang und die Komplexität der Materialbereinigung machten die Erstellung eines sehr großen Programmes erforderlich. Hier können - anhand des Ablaufplans auf der nächsten Seite - nur die wichtigsten Programmteile betrachtet werden.

61 Ablaufplan

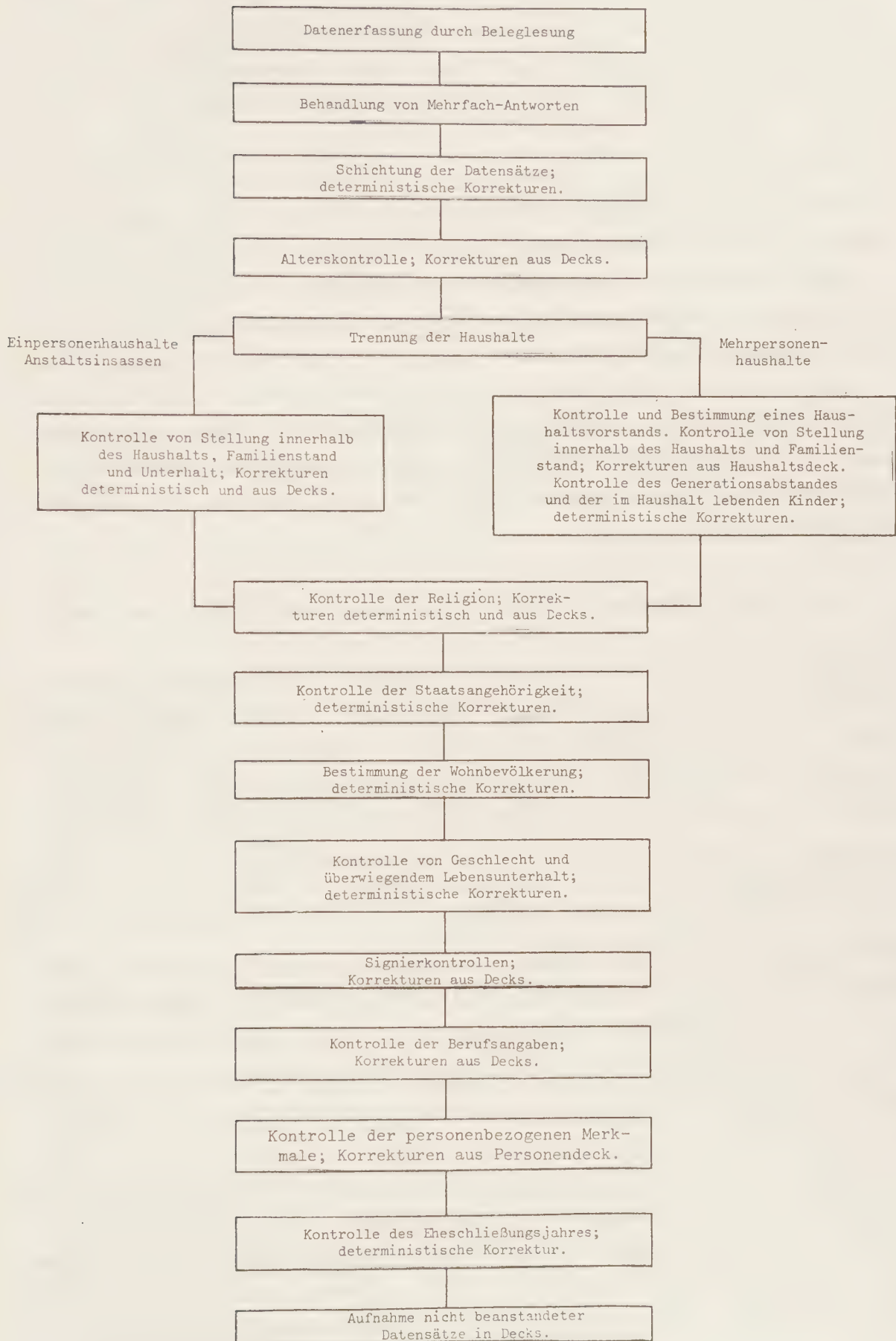
Die Volkszählung 1970 war die erste in der Bundesrepublik, bei der alle Kontrollen und Korrekturen des Fragebogens maschinell durchgeführt wurden. Allerdings waren manuelle Eingriffe in den programmierten Ablauf zugelassen, wenn sich automatische Lösungen als schlecht erwiesen. Fehlende und in bezug auf den Kontrollrahmen widersprüchliche Informationen in den Datensätzen wurden entweder deterministisch aus den Informationen eines Datensatzes bzw. den Informationen der Datensätze aller Personen eines Haushalts oder durch Übernahme von Werten nach dem Hot Deck-Verfahren bereinigt.

Durch Cold Decks war das System in einen funktionsfähigen Anfangszustand versetzt worden.¹⁾ Die Decks waren durch rd. 3 000 Datensätze "aufgeheizt" worden, um möglichst viele Deckzellen mit aktuellen Werten zu füllen.

Zunächst wurden die Datensätze jeder einzelnen Person durch Beleglesung erfaßt. Dabei wurden unzulässige Mehrfach-Antworten in den Datensätzen nach besonderen Regeln behandelt. Anschließend erfolgte die Schichtung des Materials durch Zuordnung der Datensätze zu fünf Typen. Im nächsten Schritt wurde kontrolliert,

1) Nachweis der Decks im Anhang S.176-179.

Das Fehlerbereinigungssystem bei der Volkszählung 1970



Diese Übersicht wurde aus Konzepten des Statistischen Bundesamtes, Wiesbaden, und den Programmablaufplänen des Bayerischen Statistischen Landesamtes, München, entwickelt.

ob die Typenbestimmung bei fehlendem Alter oder im Widerspruch zum Alter stattgefunden hatte. War das der Fall, dann wurden Alter und Geburtsmonat entweder aus einem Deck für die Alterseinsetzung nach Typen oder aus einem Deck für die Alterseinsetzung nach Anstaltsarten korrigiert.

Eine manuelle Prüfung dieser Einsetzungen zeigte, daß die für jede Anstaltsart geplanten Altersgrenzen angesichts der Vielfalt von Anstaltszwecken, die oft hinter einer Anstaltsart verborgen war, zu falschen Alterskorrekturen führte. Daher mußte die Spezifikation der Decks nach dem Alter in den meisten Fällen aufgegeben werden.

Da sich die nächsten Schritte der Fehlerbereinigung bei Einpersonenhaushalten und Anstaltsinsassen von denen bei Mehrpersonenhaushalten unterschieden, erfolgte nun eine Trennung der Haushalte.

Im ersten Fall standen für die Fehlerbereinigung der Merkmale "Stellung innerhalb des Haushalts", "Familienstand" und "Unterhalt" keine Informationen weiterer Haushaltsmitglieder zur Verfügung; Fehler bei der "Stellung innerhalb des Haushalts" wurden aus dem Datensatz heraus deterministisch korrigiert, Fehler bei den beiden anderen Merkmalen dagegen aus Decks bereinigt. Auch hier waren die Anstaltsdecks nach Anstaltsarten gegliedert, um die Variabilität der eingesetzten Werte sinnvoll zu beschränken.

Bei der Korrektur der haushaltsbezogenen Merkmale konnte auf Informationen der übrigen Haushaltsmitglieder zurückgegriffen werden. Die besonderen Schwierigkeiten dieser Korrekturen bestanden darin, daß sie bei den anschließend ablaufenden personenbezogenen Kontrollen erneut zu Inkonsistenzen führen konnten.

Zunächst wurde kontrolliert, ob es in jedem Haushalt einen Haushaltsvorstand und einen Ehegatten gab. Bei deutschen Ehen wurde die Altersgrenze des männlichen Ehepartners auf 18 Jahre, die des weiblichen Ehepartners auf 16 Jahre fixiert. Bei nicht-deutschen Ehen lagen die entsprechenden Grenzen bei 16 bzw. 14 Jahren. Wenn bei einer Person des Haushalts Angaben zum Familienstand und zur Stellung innerhalb des Haushalts fehlten, wurden diese Angaben aus dem Haushaltsdeck ergänzt.

Schließlich wurde der Generationsabstand der Haushaltsmitglieder mittels Nullkontrolle geprüft. Sinnvolle Einsetzungen aus dem Haushaltsdeck gelangen hier selten, weil die Korrespondenz des fehlerhaften Haushalts zu dem Musterhaushalt im Deck - besonders bei Großhaushalten, die aus mehreren Generationen bestanden - nicht hergestellt werden konnte.

Wieder gemeinsam für alle Haushalte und Anstaltsinsassen erfolgte die automatische Kontrolle und Korrektur der Angaben zur Religion. Ursprünglich war vorgesehen, die Fälle "ohne Angabe" den Gemeinschaftslosen zuzuordnen. Da aber zu vermuten war, daß die Mehrzahl der Personen, die keine Angabe zu dieser Frage machte, dies unabsichtlich tat,¹⁾ während die Gemeinschaftslosen ihre Nichtzugehörigkeit zu einer Religions- oder Weltanschauungsgemeinschaft häufig bewußt dokumentieren, wurde eine deterministische Lösung aus dem Haushaltszusammenhang für Kinder bis zu 15 Jahren vorgesehen. Wenn die Religionsangabe bei den Eltern übereinstimmte, wurde diese übernommen; ansonsten wurde aus dem Deck heraus korrigiert.

Die fehlenden Angaben zur Staatsangehörigkeit wurden deterministisch eingesetzt. Wenn bei einem Einpersonenhaushalt oder bei allen Mitgliedern eines Mehrpersonenhaushalts die Angabe fehlte, wurde "deutsch" eingesetzt. Sofern bei Mehrpersonenhaushalten "deutsch" vor "nichtdeutsch" überwog, wurde "deutsch" eingesetzt.

Für die Bestimmung der Wohnbevölkerung war die Frage nach dem weiteren Wohnsitz mit ihren Unterfragen relevant. Außerdem wurde die Schicht, in der sich der Datensatz befand sowie die Staatsangehörigkeit (deutsch oder nichtdeutsch) herangezogen. Im nächsten Abschnitt wird darauf ausführlich eingegangen.

Schließlich wurde das Geschlecht deterministisch korrigiert. Zunächst wurde kontrolliert, ob die betreffende Person "Hausfrau" oder "Soldat" markiert hatte, dann, ob Informationen über Geburtsjahre von im Haushalt lebenden ledigen Kindern vorlagen. Danach wurde über die Kontrolle des Familienstandes und des Eheschließungsjahres eine Lösung versucht. Wenn auch dies erfolglos

1) Vgl. Storch, H., Die Religionszugehörigkeit der Bevölkerung Nordrhein-Westfalens am 6. Juni 1961, in: Statistische Rundschau für das Land Nordrhein-Westfalen, Hrsg. Statistisches Landesamt Nordrhein-Westfalen, 15. Jg., H. 10, Düsseldorf 1963, S.218.

blieb, wurde alternierend "männlich"/"weiblich" eingesetzt.¹⁾
Korrekturen des Unterhalts fanden deterministisch nach Typen statt.

Es folgten automatische Signierkontrollen verschiedener Merkmale. Die Datenfelder wurden mittels Einzelfeld-Kontrollen auf Vorhandensein und Gültigkeit der Codes geprüft. Korrekturen der Signierbegriffe wurden ausschließlich aus Decks vorgenommen. Versuche, den Code plausibel zu machen, wurden nicht unternommen.²⁾

Im repräsentativen Teil der Zählung hatten die Auskunftspersonen die Religionszugehörigkeit verbal zu bezeichnen, wenn für sie die Antwortvorgaben "sonstige christliche Gemeinschaften" oder "sonstige religiöse Gemeinschaften" zutrafen. Das Kontrollprogramm sah die Löschung des Religionsschlüssels vor, wenn eine Markierung der beiden Antwortvorgaben fehlte. Es erscheint sehr fraglich, ob - selbst unter Berücksichtigung des hier vorhandenen logischen Fehlers in der Informationsstruktur - der Informationsausfall bei den "Kontrollfragen" höher zu bewerten war als die codierte Textangabe.

Die Frage nach der ausgeübten Tätigkeit/Beruf, die 1961 noch an die gesamte Bevölkerung gerichtet war, blieb 1970 auf den repräsentativen Teil der Erhebung beschränkt. Daher konnte für die weitaus größte Zahl der Fälle aus dem Bereich der Erwerbstätigkeitsmerkmale nur eine Kombinationskontrolle der Werte für "Wirtschaftszweig" und "Stellung im Beruf" durchgeführt werden. Aus den 100 codierten Wirtschaftszweigen und den 7 auf dem Fragebogen vorgegebenen Ausprägungen der beruflichen Stellung war eine Tabelle entwickelt worden, die bei 700 möglichen Feldern 97 Ausschließungen aufwies. Die Bereinigung einer ungültigen Wertekombination erfolgte durch Änderung der "Stellung im Beruf". Hier wäre im Rahmen der Fehlerallokation der signierte Begriff vermutlich als der weniger zuverlässige bestimmt worden gegenüber der Antwortvorgabe der "Stellung im Beruf". Daraus läßt sich ab-

1) Bei der rumänischen Volkszählung 1966 wurde diese Zuordnung nur für Personen unter 12 Jahren durchgeführt; Manescu, M., Electronic Data Processing of the Population and Housing Census in Romania, in: Economic Computation and Economic Cybernetics, Studies and Research, Hrsg. The Centre of Economic Computation and Economic Cybernetics, H. 4, Bucharest 1968, S.34.

2) Vgl. oben S.109.

leiten, daß die nachträgliche Überprüfung der Informationen zum Wirtschaftszweig vermieden werden sollte. Fehlende Angaben wurden aus dem Personendeck eingesetzt.

Im repräsentativen Teil wurde zuerst eine automatische Signierkontrolle des Berufsschlüssels vorgenommen, ehe die Konsistenz der Kombination dieses Merkmals mit "Wirtschaftszweig", "Stellung im Beruf", "Alter" - zur Vereinfachung reduziert auf zwei Altersgruppen - kontrolliert werden konnte. Im Falle von Rückweisungen der Kombination wurden die Merkmale einzeln aus Decks ergänzt. Hier kam es zu vielen Iterationen. Die Analyse dieses Teils der Materialbereinigung zeigt besonders deutlich, daß das empirische Wissen über die Struktur und die Fehlereigenschaften der Erwerbstätigkeitsmerkmale bisher nicht ausreicht, um gesicherte Prioritäten für einzelne Korrekturschritte zu bestimmen.

Einen Schwerpunkt im System der Fehlerbereinigung bildeten die personenbezogenen Kombinationskontrollen. Insgesamt waren, gegliedert nach Typen, 321 Kombinationsabfragen entwickelt und mit Korrekturschritten verbunden worden. Dazu ein Beispiel:¹⁾ Wenn der ausgeübte Beruf eines Erwerbstätigen "Arzt" ist, dann muß der Schulabschluß "Hochschule", die Ausbildung "Hochschule", die Ausbildungsdauer mindestens 6 Jahre, die Hauptfachrichtung "Medizin" und das Jahr des letzten Abschlusses mindestens 26 Jahre nach dem Geburtsjahr sein. Wenn eine dieser Ausprägungen fehlte oder dem Schema nicht entsprach, wurde der Wert entsprechend determiniert.

Die Zahl manueller Eingriffe in den Ablauf der Fehlerbereinigung war bei den Bildungsmerkmalen (besonders "Ausbildungsdauer" und "Hauptfachrichtung") groß. Bei der Konzeption des Systems war die Vielfalt der in der Realität vorkommenden Wertekombinationen nicht erkannt worden. Da die Informationsstruktur bei diesen Merkmalen zudem viele Lücken aufwies, kam es zu Serieneinsetzungen aus den Decks.

1) Vgl. "Wenn-Dann"-Schema im Anhang S.175.

Ähnliches galt bei der Fehlerbereinigung eines Teils der Erwerbstätigkeitsmerkmale. Hier wurden in vielen Fällen Nichterwerbstätige, die - wie z.B. Rentner - "sonstige Nebentätigkeit" markiert hatten, und Hausfrauen, die sich angesichts ihrer Hausfrauentätigkeit in der beruflichen Stellung als "mithelfende Familienangehörige" bezeichneten, durch das Programm zu Erwerbstätigen.

Nach dem Durchlauf durch alle Kontrollen erfolgte für die einwandfreien Fälle die Aufnahme in Decks. Dazu war es notwendig, die spezifischen Aufnahmebedingungen jedes Decks zu erfüllen.

62 Bestimmung der Wohnbevölkerung

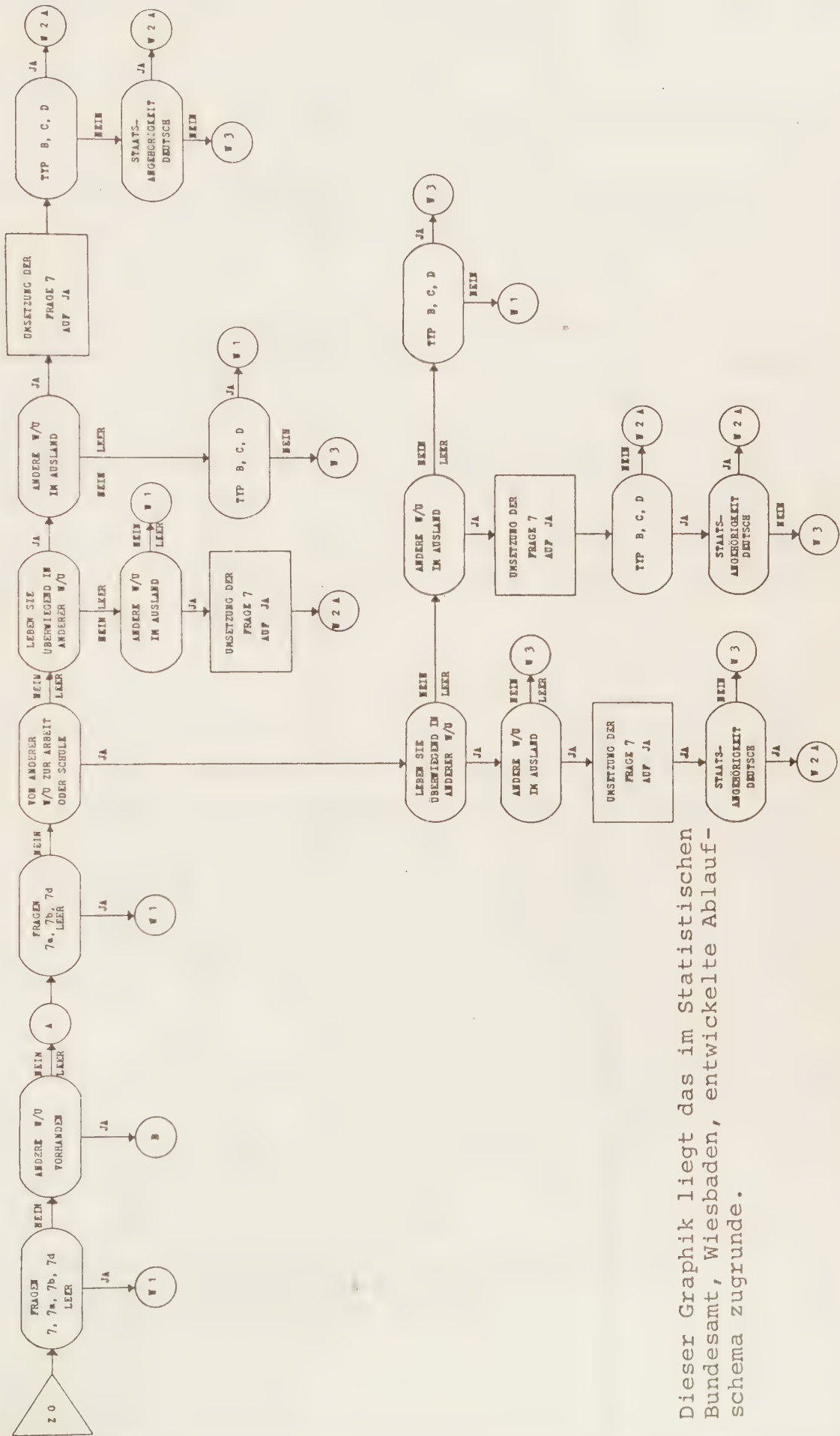
Nach der Feststellung der wohnberechtigten Bevölkerung durch die Bestandskontrollen erfolgte die Abgrenzung der Wohnbevölkerung aufgrund der Antworten zu Frage 7 des Volkszählungsbogens "Haben Sie noch eine andere Wohnung oder Unterkunft?" und ihren Unterfragen. Die Zuordnung erfolgte bei der Volkszählung 1970 erstmals maschinell.

7 Haben Sie noch eine andere Wohnung od. Unterkunft?	nein	• •
Wenn ja:	ja	—
a) Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?	ja	• •
	nein	• •
b) Leben Sie überwiegend in der anderen Wohnung/Unterkunft?	ja	—
	nein	• •
c) Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	—
	ja, mit Nebenwohn.	• •
	nicht gemeldet	• •
d) Befindet sich die andere Wohnung/Unterkunft im Ausland?	ja	• •
	nein	—

Es wurden 4 Bevölkerungsgruppen unterschieden:¹⁾

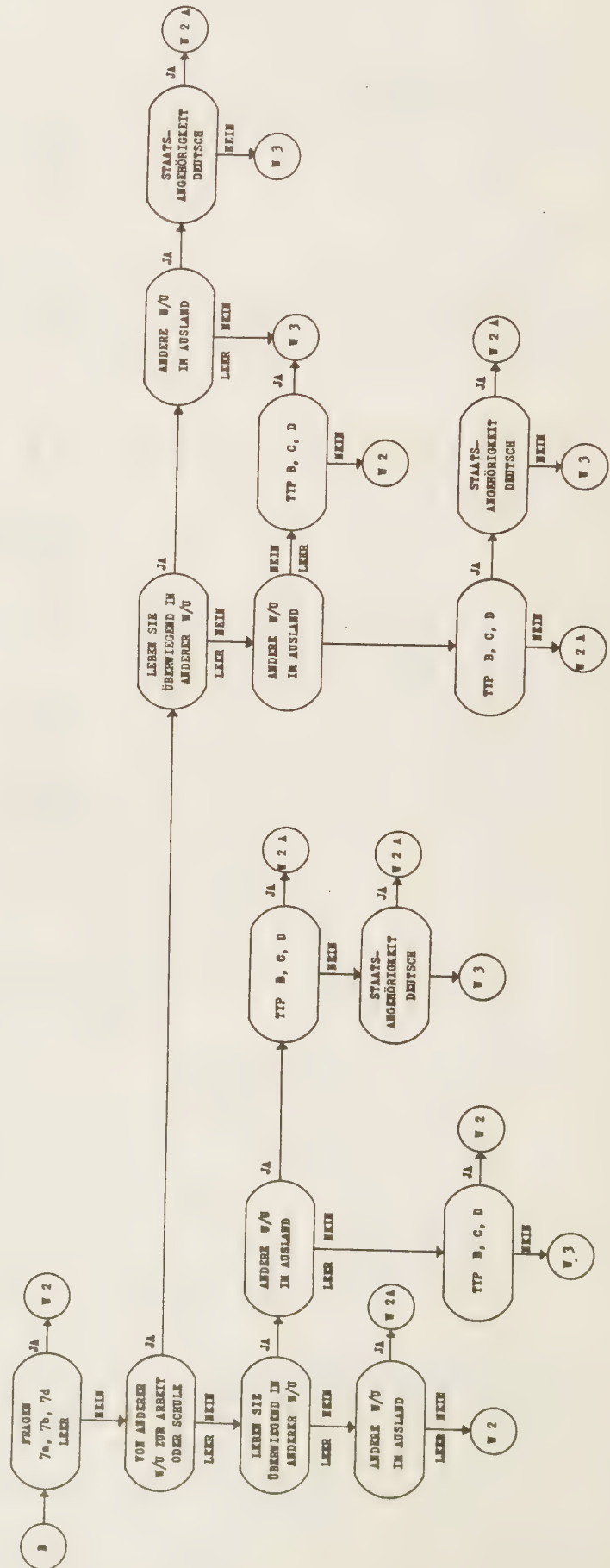
Bevölkerungsgruppe 1: Personen mit nur einer Wohnung oder Unterkunft; Wohnbevölkerung in der Zählgemeinde (W1).

1) Vgl. Programmablaufpläne auf den folgenden Seiten.



Dieser Graphik liegt das im Statistischen Bundesamt, Wiesbaden, entwickelte Ablaufschema zugrunde.

Programmablaufplan für: Bestimmung der Wohnbevölkerung bei der Volkszählung 1970



- Bevölkerungsgruppe 2: Personen mit weiterer Wohnung oder Unterkunft im Bundesgebiet; Wohnbevölkerung in der Zählgemeinde (W2).
- Bevölkerungsgruppe 2A: Personen mit weiterer Wohnung oder Unterkunft im Ausland; Wohnbevölkerung in der Zählgemeinde (W2A).
- Bevölkerungsgruppe 3: Personen mit weiterer Wohnung oder Unterkunft; Absetzung von der Wohnbevölkerung in der Zählgemeinde (W3).

Auf der Bevölkerungsgruppe 3 lag das Hauptaugenmerk. Wenn ein Datensatz diese Kennzeichnung erhielt, dann wurde die Person bei der Tabellierung als "anderswo zur Wohnbevölkerung gehörig" ausgewiesen. Die Bevölkerungsgruppen 1, 2 und 2A bildeten die "Wohnbevölkerung". Beide Gruppen sind Darstellungseinheiten, nicht wie im Falle der Typenzuordnung Aufbereitungseinheiten. Zu Zwecken der deskriptiven Fehlerforschung war nicht nur die Wohnbevölkerung sondern auch die Besetzung der Gruppen 1, 2 und 2A von Bedeutung.

Von der Eingangsfrage und den vier Unterfragen war die Frage 7c nach der behördlichen Meldung für die Bestimmung der Wohnbevölkerung nicht relevant. Die Werte wurden ungeprüft aggregiert und tabellarisch nachgewiesen. Danach hätten vier Fragen mit den möglichen Ausprägungen "ja", "nein" und "leer" entschieden werden müssen. Das wären $3^4 = 81$ Kombinationen gewesen. Da die Eingangsfrage, wenn sie keine Information enthielt, als "nein" gewertet wurde,¹⁾ reduzierte sich die Anzahl der Kombinationen auf $2 \cdot 3^3 = 54$.

Für den Fall, daß der Fragenkomplex einwandfrei beantwortet war, wurde die Zuordnung wie folgt vorgenommen (die einzelnen maschinellen Abfragen bis zur Entscheidung über die Zuordnung lassen sich im Programmablaufplan verfolgen):

1) Vermerk des Statistischen Bundesamtes, Wiesbaden, vom 12.11.1969, VII A 21, Kp/Hb: Feststellung der Wohnbevölkerung bei der Volkszählung 1970.

a) Bevölkerungsgruppe 1:

7 Haben Sie noch eine andere Wohnung od. Unterkunft?	nein	☒
Wenn ja:	ja	☐
a) Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?	ja	☐
	nein	☐
b) Leben Sie überwiegend in der anderen Wohnung/Unterkunft?	ja	☐
	nein	☐
c) Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	☐
	ja, mit Nebenwohn.	☐
	nicht gemeldet	☐
d) Befindet sich die andere Wohnung/Unterkunft im Ausland?	ja	☐
	nein	☐

b) Bevölkerungsgruppe 2:

7 Haben Sie noch eine andere Wohnung od. Unterkunft?	nein	☐
Wenn ja:	ja	☒
a) Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?	ja	☐
	nein	☒
b) Leben Sie überwiegend in der anderen Wohnung/Unterkunft?	ja	☐
	nein	☒
c) Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	☐
	ja, mit Nebenwohn.	☐
	nicht gemeldet	☐
d) Befindet sich die andere Wohnung/Unterkunft im Ausland?	ja	☐
	nein	☒

c) Bevölkerungsgruppe 2A:

7 Haben Sie noch eine andere Wohnung od. Unterkunft?	nein	☐
Wenn ja:	ja	☒
a) Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?	ja	☐
	nein	☒
b) Leben Sie überwiegend in der anderen Wohnung/Unterkunft?	ja	☐
	nein	☒
c) Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	☐
	ja, mit Nebenwohn.	☐
	nicht gemeldet	☐
d) Befindet sich die andere Wohnung/Unterkunft im Ausland?	ja	☒
	nein	☐

d) Bevölkerungsgruppe 2A bzw. 3:

7 Haben Sie noch eine andere Wohnung od. Unterkunft?	nein	☐
Wenn ja:	ja	☒
a) Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?	ja	☒
	nein	☐
b) Leben Sie überwiegend in der anderen Wohnung/Unterkunft?	ja	☒
	nein	☐
c) Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	☐
	ja, mit Nebenwohn.	☐
	nicht gemeldet	☐
d) Befindet sich die andere Wohnung/Unterkunft im Ausland?	ja	☒
	nein	☐

Ein Deutscher wurde der Bevölkerungsgruppe 2A zugerechnet, ein Ausländer der Bevölkerungsgruppe 3.

e) Bevölkerungsgruppe 2A bzw. 3:

7	Haben Sie noch eine andere Wohnung od. Unterkunft?		nein	• •
	Wenn ja:		ja	—
a)	Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?		ja	• •
			nein	—
b)	Leben Sie überwiegend in der anderen Wohnung/Unterkunft?		ja	—
			nein	• •
c)	Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	• •	
		ja, mit Nebenwohn.	• •	
		nicht gemeldet	• •	
d)	Befindet sich die andere Wohnung/Unterkunft im Ausland?		ja	—
			nein	• •

Dieser Fall war einwandfrei beantwortet, wenn es sich nicht um einen Schüler, Studierenden oder Erwerbstätigen (Typen B, C, D) handelte, der wie unter d) als Deutscher nach 2A, als Ausländer nach 3 entschieden wurde.

Für Schüler, Studierende und Erwerbstätige war es ein Fehlerfall, der 2A zugeordnet wurde.

f) Bevölkerungsgruppe 3:

7	Haben Sie noch eine andere Wohnung od. Unterkunft?		nein	• •
	Wenn ja:		ja	—
a)	Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?		ja	—
			nein	• •
b)	Leben Sie überwiegend in der anderen Wohnung/Unterkunft?		ja	—
			nein	• •
c)	Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	• •	
		ja, mit Nebenwohn.	• •	
		nicht gemeldet	• •	
d)	Befindet sich die andere Wohnung/Unterkunft im Ausland?		ja	• •
			nein	—

g) Bevölkerungsgruppe 3 bzw. 2:

7	Haben Sie noch eine andere Wohnung od. Unterkunft?		nein	• •
	Wenn ja:		ja	—
a)	Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule?		ja	• •
			nein	—
b)	Leben Sie überwiegend in der anderen Wohnung/Unterkunft?		ja	—
			nein	• •
c)	Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet?	ja, mit Hauptwohn.	• •	
		ja, mit Nebenwohn.	• •	
		nicht gemeldet	• •	
d)	Befindet sich die andere Wohnung/Unterkunft im Ausland?		ja	• •
			nein	—

Dieser Fall war einwandfrei beantwortet, wenn es sich nicht um Schüler, Studenten und Erwerbstätige handelte, die nach 3 entschieden wurden. Für diese Personengruppen war es ein Fehlerfall, der 2 zugeordnet wurde.

Außer diesen 7 genannten Antwortkombinationen waren weitere 47 im Programm. Wenn die Eingangsfrage mit "nein" markiert war, dagegen einzelne oder alle Unterfragen mit "ja", so war auch diese Inkonsistenz zu entscheiden.

Die Unterfrage 7b "Leben Sie überwiegend in der anderen Wohnung/Unterkunft?" erscheint nicht eindeutig formuliert. Zwar müßte eine Person in der Gemeinde, in der sie weitere Wohnung oder Unterkunft hat und in der sie zur Arbeit oder Ausbildung geht, überwiegend leben. Wenn sie jedoch subjektiv den Schwerpunkt ihrer Lebensbeziehungen in der Gemeinde sieht, in der sie ihre "Hauptwohnung" begründet hat, müßte das nach Meinung des Verfassers berücksichtigt werden. Das hätte dann die Zuordnung der Person zur Wohnbevölkerung dieser Gemeinde zur Folge. Nach einer Änderung des Melderechts ist vorgesehen, daß Verheiratete am Familienwohnsitz, Schüler und Studierende jedoch - wenn sie unverheiratet sind - weiterhin am Ausbildungs-ort zur Wohnbevölkerung gezählt werden.

Die Bevölkerungsgruppe 2A wurde etwas begünstigt durch die Entscheidung, die Eingangsfrage generell in "ja" umzusetzen, wenn die Zweitwohnung im Ausland bejaht wurde. Der Entscheidung lag die Prämisse zugrunde, daß den Befragten bei der Beantwortung der Eingangsfrage ein Irrtum unterlaufen sei. Diese Unterstellung war sicherlich etwas gewagt.

Bei der Analyse des Konzepts zeigt sich ein weiteres Problem. Besitzer von Ferienhäusern sind in den typischen Fremdenverkehrsgemeinden oft mit weiterem Wohnsitz gemeldet. Um diese Personen als wohnberechtigte Bevölkerung zu erfassen, wurden von diesen Gemeinden Volkszählungsbogen - wenngleich bei Abwesenheit der Personen lückenhaft - ausgefüllt. Wenn die Frage 7 bejaht wurde, die Unterfragen jedoch leer blieben, erfolgte keine Absetzung von der Wohnbevölkerung, sondern die Zuordnung zur Bevölkerungsgruppe 2; damit wurde die Wohnbevölkerung in diesen Gemeinden ungerechtfertigt erhöht. Eine nachträgliche Änderung

durch eine generelle Absetzung von der Wohnbevölkerung erschien jedoch im Hinblick auf die Struktur aller Fälle nicht ratsam.¹⁾

Das Konzept der maschinellen Zuordnung bewährte sich auch im Falle von Besuchern, wenn diese einen Fragebogen ausfüllten, obwohl sie nicht zu erheben waren. Dabei konnten Besucher ihre eigene Wohnung sogar subjektiv als "andere Wohnung oder Unterkunft" empfinden.

Deutsche, die im Ausland arbeiteten und auch überwiegend dort lebten, wurden definitionsgemäß zur Wohnbevölkerung gezählt, wenn sie bei einem Aufenthalt in Deutschland von der Volkszählung erfaßt wurden. Wenn ein Ausländer unter "anderer Wohnung oder Unterkunft" jedoch die derzeit im Bundesgebiet befindliche verstand, und er so wie in Fall d) antwortete, dann wurde er von der Wohnbevölkerung abgesetzt. Das führte zu einer unzulässigen Untererfassung.

Das Beispiel zeigt, wie intensiv alle in der Praxis vorkommenden Antwortkombinationen sowie denkbare fehlerhafte Kombinationen untersucht werden müssen. Das Programm ließ in dieser Beziehung kaum Wünsche offen. Die deterministischen Korrekturen waren problemlos bei vollständiger Informationsdichte zu diesem Fragenkomplex. Sie kamen bei abnehmender Informationsdichte natürlich nicht ohne Hypothesen aus. Hier zu Deck-Lösungen überzugehen, erwies sich jedoch nicht als opportun, da wegen der großen Bedeutung der amtlichen Feststellung der Wohnbevölkerung - vor allem für den Finanzausgleich - erwartet werden konnte, daß eine Reihe von Gemeinden das Ergebnis überprüfte. Zufallslösungen hätten hier unter Umständen weniger stichhaltige Argumente geliefert.

63 Qualitätskontrolle der automatischen Korrekturen

Die Komplexität der Fehlerstruktur und die Mängel des in der Praxis verwendeten Hot Deck-Korrekturverfahrens zeigen, daß es

1) Schreiben des Statistischen Bundesamtes, Wiesbaden, vom 19.4.1971, VIII V 4/00, Li/L/Schä: Zuordnung zur Wohnbevölkerung.

unrealistisch ist, eine vollautomatische Fehlerbereinigung - ohne manuelle Eingriffe - vorzusehen. Daher wurden in der Bundesrepublik 1970 von den automatischen Korrekturen Fehlerprotokolle erstellt, um nicht hinnehmbare Löschungen, Änderungen und Ergänzungen in Datensätzen durch manuelle "Überkorrekturen" zu bereinigen. In diesem Sinne dienen die Protokolle operativen Zwecken. Sie können aber auch - ggf. modifiziert - als Fehlerstatistiken für die deskriptive Fehlerforschung Verwendung finden.

In USA sind ebenfalls maschinelle Listungen üblich, die sowohl "corrective actions" ermöglichen als auch Indikatoren für die Qualität des Datenmaterials darstellen.¹⁾ Beim kanadischen Census 1971 hatte das Maschinenprogramm DIARY II die Aufgabe, die Auswirkungen der automatischen Fehlerbereinigung zu protokollieren. Zu diesem Zweck wurden die ursprünglich erhobenen Daten mit den bereinigten verglichen und verschiedene Kenngrößen berechnet. Diese zeigten laufend Veränderungen in der Datenstruktur an.²⁾

In der Bundesrepublik wurden Haushalte und Einzelpersonen aufgelistet. Haushalte dann, wenn bei einer Person haushaltsbezogene Merkmale automatisch korrigiert worden waren, Einzelpersonen, wenn die Untergrenze für die Einsetzung von Ersatzwerten überschritten wurde. Wie bei dem Konzept der Schichtenbildung lag auch der Bestimmung des Schwellenwertes im Protokoll die Annahme zugrunde, daß für drei fehlende bzw. widersprüchliche Datenfelder die Ersatzwerte aus der Informationsstruktur des Datensatzes abgeleitet bzw. konsistent in diese Struktur eingefügt werden könnten. Bis zu drei Korrekturen blieben die Auswirkungen der automatischen Materialbereinigung also unsichtbar.

Sicherlich ist auch hier das Kriterium der numerischen Spezifikation von Korrekturen unzureichend. Damit ist weder die Berechtigung noch der Umfang manueller Überkorrekturen zu begründen. Die Beispiele für Serieneinsetzungen haben gezeigt, daß eine Konzentration auf einzelne, vom Programm schlecht gelöste Merkmale oder Merkmalskombinationen, besser wäre.

-
- 1) Dalenius, T., Automatic Estimation of Missing Values in Censuses and Sample Surveys, in: Statistisk Tidskrift, Hrsg. Statistiska Centralbyrån, H. 7, Stockholm 1962, S.398.
 - 2) Report of the DIARY II Task Force: Specifications for the DIARY I Programme, hektographierte Arbeitsunterlage, Hrsg. Dominion Bureau of Statistics, Ottawa 1971, S.1-10.

Der Schwellenwert konnte von den Statistischen Landesämtern wahlweise höher festgesetzt werden, wenn die Qualität des auf die automatische Fehlerbereinigung treffenden Datenmaterials dies erforderte.¹⁾ In Ermangelung vorformulierter Genauigkeitsansprüche war die Übereinkunft über die Höhe des Schwellenwertes eine Ermessensentscheidung, die von Terminzwängen mitbestimmt sein konnte. Im Prinzip ist daher eine einheitliche Bewertung der Ergebnisse nicht möglich.

Zweifellos sind manuelle Überkorrekturen eine Belastung für den Arbeitsablauf, da jede Umsetzung maschinell erneut auf Konsistenz geprüft werden muß. Dennoch sind sie zur Qualitätsverbesserung unumgänglich.

Die bei der Volkszählung 1970 erstmals erstellten Fehlerstatistiken, die alle Operationen der automatischen Fehlerbereinigung enthielten, sind von unschätzbarem Wert für Fehleranalysen und stellen eine wichtige Grundlage für die Planung der Fehlerbereinigung der nächsten Zählung dar. Auf die Auswertungen durch das Statistische Bundesamt darf man gespannt sein.

Eine der Tabellen, die von Bayern zu Testzwecken ausgedruckt worden war, hatte folgendes Aussehen:

Korrekturart	Fälle	%
Mehrfach-Antworten	26 935	28
deterministisch personenbezogen	24 957	26
deterministisch haushaltsbezogen	15 008	16
aus dem Personendeck	25 670	27
aus dem Haushaltsdeck	794	1
aus anderen Decks	2 420	2
insgesamt	95 784	100

Diese Tabelle liegt in regionaler Gliederung bis herunter zu Häuserblöcken bzw. Straßenseiten vor. Sie soll hier lediglich zur Illustration dienen, da sie wohl nicht als repräsentativ angesehen werden kann.

1) In Nordrhein-Westfalen wurde der Schwellenwert bei vier Fehlern festgelegt. In der rumänischen Volkszählung waren ebenfalls vier Fehler vorgesehen; Manescu, M., Electronic Data Processing of the Population and Housing Census in Romania, a.a.O., S.33.

Weitere ausgezeichnete analytische Möglichkeiten bieten die Aufzeichnungen der Datensätze im Originalzustand und im Zustand nach der Fehlerbereinigung. Damit können z.B. die Auswirkungen der Abschneidegrenzen auf die Häufigkeitsverteilungen untersucht werden. Es ist zu hoffen, daß die Finanzierung wenigstens einiger Forschungsprojekte, die auf der Basis dieses Datenmaterials möglich sind, gelingt.

7 Bewertung der automatischen Fehlerbereinigung

Wenn bei den Daten der Volkszählung nicht - wie bei quantitativen Merkmalen - die Möglichkeit gegeben ist, extreme Werte in univariaten oder multivariaten Verteilungen zu löschen oder durch solche zu ersetzen, so daß die Schätzung der Parameter durch den Analytiker - in welchem Sinn auch immer - nicht mehr "gestört" ist, wenn es keine Kovarianzmatrix gibt, an der sich der Erfolg der Bemühungen ablesen ließe, dann werden andere Kriterien benötigt, zum Beispiel die, daß kein Wert mehr zu Kontrollverletzungen führt. Genügt das nicht, dann können die Ergebnisse mit Kontrollerhebungen verglichen werden, die - wenn sie auch keine "wahren" Werte bringen -, so bei entsprechender Sorgfalt doch "optimale". Als eine der Voraussetzungen für diese Sorgfalt wird immer wieder die manuelle Aufbereitung bzw. Fehlerbereinigung angesehen. Darunter kann im Grunde nur die Verifikation einer beanstandeten Information bei der Auskunftsperson gemeint sein. Ist das für eine Großzählung realistisch? Und wenn aus technisch-organisatorischen Zwängen heraus Zuflucht zu automatischen Verfahren genommen wird, was ist zu tun, um ihre negativen Auswirkungen einzudämmen für den Fall, daß mit ihnen die gewünschte Genauigkeit nicht zu erzielen ist?

71 Vorteile gegenüber manuellen Verfahren

Es stehen nicht allzu viele Informationen über Fehleranteile in statistischen Daten zur Verfügung, die Aufschluß geben könnten über den Arbeitsumfang bei der Fehlerbereinigung. Die wenigen sind zudem heterogen.

Terry kommt auf Grund seiner langjährigen Fehlerbereinigungspraxis mit Daten aus dem Ingenieurwesen, aus den Naturwissenschaften und wirtschaftlichen Erhebungen zu dem Schluß, daß zwischen 0,1 und 45 % der Datensätze zu Beanstandungen geführt hatten, wobei er den mittleren Fehleranteil auf 5 % schätzt.¹⁾

1) Terry, M. E., The Principles of Statistical Analysis Using Large Electronic Computers, in: Bulletin of the International Statistical Institute, Proceedings of the 34th Session, Vol.40, Book 1, Ottawa 1963, S.549.

Bei der Volkszählung 1961 in der Bundesrepublik waren durchschnittlich 3-5 % der Wohnbevölkerung (das entspricht Datensätzen) als fehlerhaft zurückgewiesen worden.¹⁾

Einem schweizerischen Beitrag ist zu entnehmen, daß auf rd. 15 % der Erhebungspapiere einer Volkszählung mindestens eine Frage nicht oder falsch beantwortet worden war.²⁾

Bei der auf S.131 dargestellten Übersicht aus der Volkszählung 1970 bezogen sich die 95 784 Korrekturen auf eine Arbeitsmasse von 73 809 Datensätzen. 59 % dieser Datensätze waren dabei durch die automatischen Kontrollen reklamiert worden.

Die erheblich voneinander abweichenden Anteile müssen nicht Ausdruck großer Qualitätsunterschiede sein. Sie dürften vielmehr auf ungleichen Voraussetzungen und Meßverfahren beruhen. 1970 war jede maschinelle Operation aufgezeichnet worden, so z.B. auch die automatische Einsetzung der Information "nicht erwerbstätig" bei Kleinkindern, die von vielen Auskunftspersonen vermutlich weggelassen worden war, weil die Kombination als selbstverständlich galt. Wenn die Fehlerstatistik außerdem auf Datenmaterial aus dem repräsentativen Teil der Zählung beruhte, war die Gefahr teilweiser Informationsausfälle viel größer. Der schweizerischen Studie lag offenbar eine Sichtkontrolle zugrunde. Der Verfasser hält aus seinen eigenen bei der Volkszählung 1970 gewonnenen Erfahrungen bei den von den Gemeinden angelieferten Belegen Fehlerquoten von 15 bis 20 % für realistisch. Der geringe bei der Volkszählung 1961 nachgewiesene Anteilsatz repräsentiert die Qualität des Datenmaterials vor der automatischen Fehlerbereinigung, nachdem intensive manuelle Bearbeitung vorausgegangen war.

Die Zahl der Kontrollabfragen wächst mit dem Frageprogramm. Sie kann aber auch bei konstantem Frageprogramm wesentlich an Umfang zunehmen, wenn die Kontrollkriterien vermehrt und die Kontroll-

1) Vgl. Diskussion über den Beitrag von Szameitat, K. und Zindler, H. J., The Reduction of Errors in Statistics by Automatic Corrections, a.a.O., S.466.

2) Bugmann, A., Statistische Techniken und Datenverarbeitung bei Großzählungen, in: Schweizerische Zeitschrift für Volkswirtschaft und Statistik, Hrsg. Schweizerische Gesellschaft für Statistik und Volkswirtschaft, 107. Jg., H. 3, Bern 1971, S.662.

grenzen verkleinert werden. Die hohen Verarbeitungsgeschwindigkeiten moderner Datenverarbeitungsanlagen begünstigen eine solche Entwicklung. Zudem können Spezialisten mit großer Projekterfahrung Konzepte erarbeiten, die maschinell strikt und konsistent durchgeführt werden. Manuelle Kontrollen müssen dagegen mit einem Stab von Mitarbeitern durchgeführt werden, in dem sich die einzelnen durch ihre geistigen Fähigkeiten, Lebenserfahrungen, Arbeitseinstellung und physische Leistungsfähigkeit unterscheiden. Lernprozesse im Zeitablauf haben zur Folge, daß die Qualität älterer Arbeitsmassen schlechter ist. Dieselben Argumente gelten für die automatische Fehlerbeseitigung. Korrekturen manuell zu erledigen, wenn die Kontrollen maschinell erfolgten - wie das 1961 in der Bundesrepublik der Fall war - kann unter arbeitsorganisatorischen Gesichtspunkten nur eine Übergangslösung darstellen. Der Rationalisierungseffekt würde gefährdet, wenn auf eine Integration von Fehleridentifikation und Korrekturen in einem umfassenden Fehlerbereinigungsprogramm verzichtet würde.¹⁾

Es mag verschiedene Gründe gegeben haben dafür, daß die amtliche Statistik in der Bundesrepublik auf automatische Korrekturen bei der Volkszählung 1961 verzichtete. Die damals eingesetzten Computer, die noch Lochkarten verarbeiteten, waren für diese komplexe und aufwendige Arbeit weitgehend ungeeignet. Andererseits ist bei der Einführung moderner Verfahren (z.B. Stichprobentechniken) eine gewisse Vorsicht zu beobachten. Sie erscheint dem Verfasser im Hinblick auf die qualitativen Auswirkungen automatischer Korrekturen auch nicht unberechtigt.

72 Grenzen der Anwendbarkeit

Der Vorteil der Konsistenz in der Beachtung von Kontroll- und Korrekturregeln kann dann zum Nachteil werden, wenn die Fehler-

1) Vgl. Fellegi, I. P., Goldberg S. A., Some Aspects of the Impact of the Computer on Official Statistics, in: Bulletin of the International Statistical Institute, Proceedings of the 37th Session, Vol. XLIII, Book 1, London 1969, S.159.

struktur falsch analysiert wird und es zur konsistenten Wiederholung logischer Fehler bei Abfragen und bei der Einsetzung der Ersatzwerte kommt. Hier könnten manuelle Arbeiten vieler Beteiligter Ausgleichswirkungen erzielen. Die Hypothese eines Fachmannes, die ihren Niederschlag in der Formulierung der Kontroll- bzw. Korrekturregel im Maschinenprogramm findet, führt oft zur Bestätigung dieser Hypothese, auch wenn sie sich als nicht haltbar erweist. Lernprozesse in manuellen Abläufen können selbst unter Berücksichtigung schlechterer Qualität von bereits erledigten Arbeitsmassen günstiger sein als die Inflexibilität eines einmal programmierten Fehlerbereinigungssystems.

Zindler schließt aus seinen Untersuchungen, daß die Korrekturverfahren relativ geringfügige Auswirkungen auf die statistischen Ergebnisse haben. "Damit würde man sich im Bereich der amtlichen Statistik doch ermutigt fühlen, Verfahren als auch sachlich sinnvolle Korrekturmöglichkeiten zu benutzen, die vom rein technischen Standpunkt aus ohnehin notwendig sind".¹⁾

Unzweifelhaft sind Zindlers Modellannahmen hinsichtlich der Entstehung der Fehler nur im Aufbereitungsprozeß, der Zufallseigenschaften der Fehlerentstehung und der schichtenweisen Vorsortierung des Materials vor der Korrektur zu streng.²⁾

Es ist eine technisch-organisatorische Notwendigkeit, die Datensätze in regionaler Ordnung zu speichern. Dadurch wird der Vorteil, den die Anpassung der automatischen Korrekturen an regionale Strukturen mit sich bringen kann, überkompensiert durch systematische Fehler, die in jeder Phase des Zählungsablaufes entstehen. Bei der Volkszählung 1970 in Nordrhein-Westfalen hat sich deutlich gezeigt, daß in bestimmten Bereichen einzelne Merkmalsausprägungen oder Kombinationen von Ausprägungen serienweise als Korrekturgrundlage für Fehlerfälle dienten. Diese Verfälschungen des Materials konnten nicht hingenommen werden, so daß die Option für manuelle Überkorrekturen

1) Zindler, H. J., Die Auswirkungen maschineller Korrekturverfahren, a.a.O., S.228/229.

2) So auch Sunter, A. B., Patrick, C. A., Binder, D. A., On the Editing of Survey Data, Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975, S.2.

ausgenutzt wurde. Immerhin wurden 5 % der gesamten Aufbereitungszeit dafür verbraucht.

Die einzige Studie über den Zusammenhang zwischen Aktualität und Qualität einer statistischen Erhebung wurde bisher von Nordbotten durchgeführt.¹⁾ Um eine Reihe von Alternativen berücksichtigen zu können, die praktisch Variationen des gesamten Ablaufs einer Zählung von der Datensammlung über die einzelnen Phasen der Aufbereitung bis hin zur Tabellierung darstellen, hat er zur Nachbildung der komplexen Zusammenhänge die Monte Carlo-Simulation gewählt. Es wurden realistische Daten aus der in Norwegen jährlich stattfindenden Erhebung im Bergbau und in der Grundstoffindustrie mit simulierten Fehlern zugrundegelegt. Die Kontrollverfahren waren Quotienten-Kontrollen mit unterschiedlichen Kontrollbereichen. Werte, die aus sorgfältig durchgeführten Nacherhebungen stammten, wurden als "wahre" Werte eingesetzt. Als automatisches Korrekturverfahren wurde das Hot Deck-Verfahren analysiert. Die Ergebnisse wurden in unterschiedlicher sachlicher und regionaler Tiefe nachgewiesen. Die Alternativen wurden durch Verwendung von "Qualitätsindikatoren" verglichen. Die besten Ergebnisse wurden erzielt beim Einsatz automatischer Kontrollen mit relativ breiten Toleranzgrenzen und Nacherhebung der zurückgewiesenen Fälle. Die Auswirkungen des Hot Deck-Verfahrens erwiesen sich als unbefriedigend.

Wenn also einerseits systematische Fehler bei automatischen Korrekturen in Kauf genommen werden müssen, andererseits aus Kosten- und Aktualitätsgründen Rückfragen bei den Auskunftspersonen nicht in Frage kommen, gilt es, einen Standard über die Größe des Fehleranteils zu formulieren, dessen Beseitigung durch automatische Korrekturen zu vertretbaren Qualitätseinbußen führt. Dabei haben Volkszählungen den Vorzug, etwa im Vergleich zu Unternehmensstatistiken, daß ein einzelner Datensatz nur geringfügige Auswirkungen auf die Ergebnisse hat. Wenn - wie von

1) Nordbotten, S., The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means for Improving the Quality of Statistics, S.432-435. Bereits früher hatte Nordbotten ein stark vereinfachtes Modell zur Ermittlung der kostenminimalen Faktorkombination bei gegebenen Qualitätsanforderungen entwickelt; Nordbotten, S., On Errors and Optimal Allocation in a Census, in: Skandinavisk Aktuarietidskrift, 40.Jg., H. 1, Uppsala 1957, S.1-10.

der Regionalplanung nicht selten gefordert - der Nachweis von Volkszählungsergebnissen zum Einzelnachweis ausartet, dann verbieten sich stochastische Korrekturen von selbst; aber auch deterministische Korrekturen wären, abgesehen von wenigen Ausnahmen, unbrauchbar. Das ist jedoch nicht ein Problem automatischer Korrekturen, sondern von Korrekturen bei Verzicht auf Nacherhebungen schlechthin.

Nun läßt sich die Grenzfehlerrate nach wissenschaftlichen Kriterien nicht bestimmen. Sie müßte - unter Festlegung eines Intervalls - von Produzenten und Konsumenten vereinbart werden. Es wurde bereits oben darauf hingewiesen, daß das bei Volkszählungen praktisch unmöglich ist. Es bleibt daher eine gewisse Subjektivität der Statistischen Ämter. In USA gilt ein automatisch korrigierter Anteil von 1-2 % an den statistischen Ergebnissen als Zielvorstellung. Wenn die Anforderungen an die Aktualität groß sind, werden auch 5 % für annehmbar gehalten. Als noch besser für wichtige Ergebnisse aus Totalzählungen wird es angesehen, wenn der auf automatische Korrekturen entfallende systematische Fehler höchstens die Hälfte des Gesamtfehlers (Wurzel aus Mean Square Error) der statistischen Ergebnisse mit allen Zufalls- und systematischen Fehlern (ohne die aus automatischen Korrekturen herrührende Fehlerkomponente) ausmacht. Dieser Wert wird, soweit er sich mit Hilfe deskriptiver Kontrollen überprüfen läßt, in USA auf Bundesebene erreicht.¹⁾

Mit automatischen Korrekturen wird weder eine nicht vorhandene Perfektion vorgetäuscht, noch handelt es sich dabei um einen Beweis für das Selbstverständnis der amtlichen Statistik.²⁾

Die Schätzung von Ergebnissen in Tabellenfeldern "ohne Angabe" durch den Konsumenten ist fehleranfälliger, als wenn die amtliche Statistik unter Zuhilfenahme des theoretischen Instrumen-

1) Pritzker, L., Ogus, J., Hansen, M. H., Computer Editing Methods. Some Applications and Results, in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.445/446.

2) Weichselberger, K., Genauigkeitsansprüche, Fehler und Kontrollen bei Volkszählungen, in: Allgemeines Statistisches Archiv, Bd 54, H. 1, Göttingen 1970, S.2.

tariums, praktischer Erfahrungen und der Möglichkeiten, die elektronische Datenverarbeitungsanlagen bieten, Ergänzungen und Bereinigungen selbst vornimmt, um den Informationswert des vorhandenen Materials zu verbessern.¹⁾ Welche Rolle die Fehlerforschung dabei spielt, soll im folgenden untersucht werden.

1) "Automatic correcting methods will only help to exploit effectively the information provided by the data which have already been collected"; Nordbotten, S., Automatic Editing of Individual Statistical Observations, a.a.O., S.11.

8 Bedeutung der Fehlerforschung

Die zur Durchführung einer Zählung zur Verfügung stehenden Ressourcen sollten sinnvollerweise nur zum Teil der Verbesserung des Konzepts der Fehlerbereinigung dienen; wichtiger erscheint, das gesamte Zählungswerk daraufhin zu untersuchen, wo in einzelnen Arbeitsphasen spürbare Qualitätsverbesserungen notwendig und möglich sind. Mit Hilfe deskriptiver Kontrollen einzelner Fehlerarten kann die Genauigkeit von Arbeitsgängen gemessen werden. Die Bewertung der Zählungsergebnisse gibt dem Konsumenten dagegen Hinweise darauf, wieviel Vertrauen er den Ergebnissen entgegenbringen darf. Die neuere Fehlerforschung bei Total- und Stichprobenerhebungen konzentriert sich auf den Nichtstichprobenfehler. Es wird zwar eingeräumt, daß die Stichprobentheorie Gutes geleistet hat zur Begründung der Bestimmung statistischer Einheiten in einer Auswahl und auch bei der Ermittlung von Schätzwerten aus Stichprobenbeobachtungen unter der Prämisse exakter Messung oder Beobachtung. Um aber Qualitätsverbesserungen, Kosteneinsparungen und Aktualitätsgewinne zu erzielen, ist eine Neuorientierung der Forschungsarbeiten erforderlich.¹⁾

81 Erfassungsfehler

Eines der größten Probleme bei Volkszählungen ist das der Unter- oder Übererfassung der statistischen Gesamtheit. Die deskriptiven Kontrollen zur Durchleuchtung der Erfassungsgenauigkeit sind in der Bundesrepublik und in den USA vergleichbar. In der Bundesrepublik wird unterschieden zwischen Sofortkontrollen sowie "Alphabetkontrollen" (1961) und "Geburtstagskontrollen" (1970). Sofortkontrollen werden 4 - 6 Wochen nach dem Zählungsstichtag durchgeführt. Sie sollen Aufschluß über die topographische Genauigkeit der Zählung bringen. Dazu ist eine Zusammenführung der Erhebungsunterlagen der Kontrollerhebungen mit denen der Haupte-
rhaupterhebungen

1) Hansen, M. H., Waksberg, J., Research on Non-Sampling Errors in Census and Surveys, a.a.O., S.317.

bung erforderlich.

Hierbei treten zwei Probleme auf: Es ist nicht völlig auszuschließen, daß wegen der zeitlichen Nähe der Kontrollerhebung zur Haupterhebung etwa entdeckte Untererfassungen operativ genutzt werden, wodurch dann das Ergebnis in Frage gestellt wird und daß vor allem solche Personen, die in der Haupterhebung schwer festzustellen sind, in der Kontrollerhebung ebenso wenig entdeckt werden können. Das gilt besonders für fehlende Personen in Haushalten.

Neben der Auswertung der Ergebnisse solcher Nacherhebungen wurde in Kanada eine analytische Methode verwendet,¹⁾ bei der die Bevölkerung am Zählungstichtag nach Alter und Geschlecht geschätzt wurde. Ausgehend von der letzten Volkszählung wurden Geburten sowie Sterbefälle und die Wanderungsbewegung registriert. Die Methode hat noch Mängel, weil das Problem der Auslandswanderung ungelöst ist und gute Schätzungen für fehlende Personen in der letzten Volkszählung nicht vorliegen. Abgesehen davon, daß außer Alter und Geschlecht kein weiteres Merkmal einbezogen werden kann, ist auch nachteilig, daß keine regionale Differenzierung möglich ist.

Durch Stichprobenkontrollen (Auswahlgrundlage war bei der Volkszählung 1961 der Anfangsbuchstabe "A" des Familiennamens, bei der Volkszählung 1970 der 31. eines jeden Monats als Geburtstag) wurde in der Bundesrepublik die korrekte Erfassung der wohnberechtigten Bevölkerung sowie die Reduktion zur Wohnbevölkerung geprüft. Bei der Zählung 1961 war diese Kontrolle als operative Kontrolle geplant. Da der Zeitaufwand jedoch starke Verzögerungen beim Nachweis erster Ergebnisse erwarten ließ, mußte dieses Vorhaben aufgegeben werden.²⁾

-
- 1) Fellegi, I. P., The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences, in: International Statistical Review, Vol. 41, No. 1, Edinburgh 1973, S.3.
 - 2) Herberger, L., Verfahrenskontrollen zur Prüfung der Vollständigkeit und Genauigkeit bei Volks- und Berufszählungen, in: Allgemeines Statistisches Archiv, Bd 54, H. 1, Göttingen 1970, S.62/63.

Nach Ausschaltung der sich in beiden Kontrollen überdeckenden Fehler wurde 1961 eine Untererfassung von 0,4 bis 0,6 % der Wohnbevölkerung festgestellt.¹⁾ Im Vergleich dazu wies der US Census of Population and Housing 1960 eine Untererfassung von rd. 3 % auf.²⁾ In Kanada waren es ebenfalls 2 bis 3 %, wobei die schwerer zu erfassende Gruppe von Männern im Alter zwischen 20 und 24 Jahren fast 10 % ausmachte.³⁾

Im Abschnitt über die Bestandskontrollen hat der Verfasser darauf hingewiesen, daß die Schnelligkeit der maschinellen Verarbeitung des Datenmaterials dazu genutzt werden kann, frühzeitig Abweichungen zu den erwarteten Sollwerten zu protokollieren. Es können daher im Rahmen der Zählungsorganisation Vorkehrungen gegen Untererfassungen getroffen werden.

82 Totaler Informationsausfall

In einer amerikanischen Studie⁴⁾ wurden die Auswirkungen der Hot Deck-Korrekturen auf verschiedene demographische und sozio-ökonomische Merkmale bei den im Erhebungszeitraum nicht betroffenen Haushalten verglichen mit den als "wahre" Werte angenommenen Ergebnissen auf Grund von Nacherhebungen. Verweigerer wurden nicht einbezogen mit der Begründung, für ihr Verhalten seien "persönliche" Gründe maßgebend. Das ist ein Nachteil der Untersuchung; von den 4,4 % ausgefallenen Haushalten waren etwa ein Drittel Verweigerer.

-
- 1) Herberger, L., Verfahrenskontrollen zur Prüfung der Vollständigkeit und Genauigkeit bei Volks- und Berufszählungen, a.a.O., S.65.
 - 2) Waksberg, J., Pritzker, L., Changes in Census Methods, in: Journal of the American Statistical Association, Vol. 64, No. 328, Menasha, Wisc. 1969, S.1146.
 - 3) Fellegi, I. P., The Evaluation of the Accuracy of Survey Results, a.a.O., S.2.
 - 4) Palmer, S., Jones, Ch., A Look at Alternate Imputation Procedures for CPS Noninterviews, hektographierte Veröffentlichung, Hrsg. US Bureau of the Census (Statistical Methods Division), Washington, D.C. 1967.

Die automatisch korrigierten statistischen Ergebnisse für diese Non Response-Fälle zeigten in den Hauptgruppen "Erwerbstätige", "Nichterwerbstätige" und "Arbeitslose" keine auffallenden Unterschiede zu den Nacherhebungen. Aber die kleinere Gruppe der "Arbeitsuchenden" war unter den als Korrekturpotential dienenden vorhandenen Informationen nur mit einem Anteil von 2,3 % vertreten, während ihr Anteil an den Nacherhebungen 7,6 % betragen hatte. Weitere Unterschiede waren: 14 % Einpersonenhaushalte unter den vorhandenen Haushalten zu rd. 33 % unter den nacherhobenen; durchschnittliche Haushaltsgröße (nur Personen über 14 Jahre) 2,3 zu 1,8. Das mittlere Alter der nacherhobenen Non Response-Haushalte lag um 5 Jahre höher als das mittlere Alter der vorhandenen Haushalte. Das jährliche Einkommen der Non Response-Haushalte war um mehr als 1000 Dollar niedriger als das der anderen Haushalte. Die Autoren hatten die Non Response-Haushalte in 4 Gruppen gegliedert: "kurze Zeit abwesend", "längere Zeit abwesend", "Verweigerer" (Informationen wurden aus älteren oder vergleichbaren Erhebungen gewonnen), "aus verschiedenen Gründen unerreichbar". Diese Gruppen wiesen größere Verschiedenheit untereinander auf als zusammengekommen von den vorhandenen Haushalten. Somit zeigt diese Untersuchung, daß totale Informationsausfälle aus den Verteilungen der zeitgerecht erhobenen Merkmale nicht genau genug korrigiert werden können.

83 Antwortfehler

Wichtige Erkenntnisse über den Antwortfehler lassen sich aus Merkmalskontrollen gewinnen. Das in USA seit 1950 im Anschluß an den Census durchgeführte Auswertungsprogramm (Evaluation and Research Program of the Censuses of Population and Housing) enthält eine Reihe solcher Kontrollen. Ausgewählte Merkmale bestimmter Personen aus der Volkszählung werden dabei mit Nacherhebungen, der zeitlich nächsten laufenden Bevölkerungserhebung oder anderen Erhebungen mit denselben Merkmalen bzw. mit Merkmalen, die in der Verwaltung anfallen, verglichen. Es werden mehrere Kenngrößen berechnet, z.B. Gross Difference Rates; das sind Anteile von Fällen, in denen die Kontrollerhebung von der Haupterhebung

abweicht. Damit läßt sich der Antwortfehler und seine Komponenten, die Antwortvarianz und die systematische Komponente, abschätzen.

Ein anderes Verfahren ist das der "interpenetrating samplings". Angenommen, in einem Gebiet soll die Anzahl der Personen mit Hochschulausbildung geschätzt werden. Dann werden zwei Schätzungen für dasselbe Gebiet durchgeführt. Das geschieht dadurch, daß die eine Hälfte zufällig ausgewählter Personen einer Gruppe von Interviewern, die andere einer zweiten zugeteilt wird. Der Unterschied zwischen beiden im Anschluß an die Erhebung vorgenommenen Schätzungen spiegelt beide Ursachen für die Varianz wider, die Antwortvarianz zwischen beiden Schätzungen zur selben Zeit unter denselben Bedingungen bei verschiedenen Interviewern, die Stichprobenvarianz zwischen beiden Stichproben zur selben Zeit. Die Stichprobenvarianz kann nun isoliert geschätzt werden und so - durch Subtraktion - ein Schätzwert für die Antwortvarianz erhalten werden.

Etwa seit 1950 wurde in amerikanischen Untersuchungen den Auswirkungen der Interviewertätigkeit große Aufmerksamkeit gewidmet.¹⁾ Bei vielen Merkmalen war der Einfluß der Interviewer auf den Antwortfehler als ausschlaggebend für Genauigkeitsverluste der Statistiken in kleinräumiger Gliederung erkannt worden. Das führte zu Modifikationen im Zählungskonzept. Das Interview wurde durch die Selbsteintragung der Auskunftspersonen ersetzt. Die Erfahrung, daß selbst bis herunter zu Gebieten mit 2000 Einwohnern der Stichprobenfehler vernachlässigbar ist im Vergleich zu systematischen Fehlerkomponenten bewirkte, daß heute fast alle Merkmale mit einem Auswahlatz von 20 %, einige sogar mit einem solchen von 15 % erhoben werden. Die Kosteneinsparungen und der Gewinn an Aktualität sind offensichtlich. Unmittelbare Konsequenz für die Erhebungsphase ist, daß ein größerer Zählerstab auf schwer zu erfassende Gebiete konzentriert werden kann. Zweifellos ist bei Anwendung der Stichprobenverfahren der Nachweis von Ergebnissen

1) Vgl. Hansen, M. H., Hurwitz, W. N., Marks, E. S., Mauldin, W. P., Response Errors in Surveys, in: Journal of the American Statistical Association, Vol. 46, No. 254, Menasha, Wisc. 1951, S.147-190; Hansen, M. H., Hurwitz, W. N., Bershad, M. A., Measurement Errors in Censuses and Surveys, in: Bulletin of the International Statistical Institute, Proceedings of the 32nd Session, Vol. XXXVIII, Book 2, Tokyo 1961, S.359-374.

für kleine Wohnplätze und Wohnblöcke wie in der Bundesrepublik nicht mehr möglich, aber leider gibt es hier keine Untersuchungen darüber, wie zuverlässig derart kleinräumige Ergebnisse sind.

84 Lesefehler

Auch wenn die erstmals 1970 eingesetzten optischen Lesegeräte noch keinen spürbaren Gewinn für die Aktualität der Ergebnisse brachten, war mit ihrem Einsatz doch ein beträchtlicher Rationalisierungserfolg verbunden. Es wird geschätzt, daß die Beleglesung im Bundesgebiet rd. 500 Datentypistinnen substituiert hat.¹⁾ Mit diesen Geräten konnten Strichmarkierungen, Maschinschrift sowie handschriftliche Ziffern erkannt werden. Diese Art der Datenerfassung war aber gleichzeitig eine Quelle für bisher nicht bekannte Fehler: Lesefehler. Dieser Fehlertyp läßt sich unterscheiden in Strichmarkierungs-Fehler, Handschrift-Lesefehler und Substitutionen als besondere Form der Handschrift-Lesefehler.

Strichmarkierungs-Fehler treten auf, wenn die "Dichte der Schwarzinformation" einzelner Zeichen nicht ausreicht sowie die Begrenzungswannen verfehlt oder überschritten werden.

Wenn der Lese-prozeß kontrolliert werden soll dadurch, daß Belege mit einer Mindestzahl an Strichmarkierungs-Fehlern zurückgewiesen werden, dann besteht die Schwierigkeit, technische Fehler von erhebungsbedingten zu unterscheiden.²⁾ Während bei technischen Mängeln die Belege oft ohne fachbezogene Korrekturen erneut eingelesen werden können, weisen die anderen Informationsausfälle auf, die der planmäßigen automatischen Fehlerbereinigung zugeführt werden müßten. Das Kontrollprogramm 1970 sah u.a. Rückwei-

1) Deininger, R., Maschinelle Beleglesung bei der Volkszählung 1970 in der Bundesrepublik Deutschland, in: IBM Nachrichten, 22. Jg., H. 210, Verlag IBM Deutschland, Sindelfingen 1972, S.126.

2) In USA bestand ein technischer Fehler darin, daß ein geschwärzter Kreis auf der Vorderseite eines FOSDIC-Belegs auf die Rückseite des vorherigen abfärbte und dort eine gültige Information erzeugte, wodurch die Altersverteilung der in USA lebenden Inder empfindlich gestört wurde; Akers, D. S., Larmon, E. A., Indians and Smudges on the Census Schedule, in: Proceedings of the Social Statistics Section, Hrsg. American Statistical Association, Washington, D.C. 1967, S.369-373.

sungen vor, wenn drei der sechs für die Fehlerbereinigung wesentlichen Merkmale "Geschlecht", "Familienstand", "Stellung innerhalb des Haushalts", "Zugehörigkeit zur Wohnbevölkerung", "überwiegender Lebensunterhalt" und "Erwerbstätigkeit" nicht zu identifizieren waren.¹⁾ Das Erfordernis manueller Korrekturen in dieser Phase des Ablaufs resultiert im Prinzip aus einem Mangel an Vertrauen in das Fehlerbereinigungsprogramm. Natürlich handelt es sich bei diesen Belegen um besondere Problemfälle, die bei der Schichtenbildung durch Zwangszuordnung hätten gelöst werden müssen. Wenn das Konzept der automatischen Fehlerbereinigung die vorhandene Informationsstruktur jedoch umfassend nutzt, sind manuelle Einsetzungen an dieser Stelle überflüssig.

Der Handschrift-Lesefehler kam dadurch zustande, daß Ziffern entweder zu schwach geschrieben worden waren, daß ihre Kontur Unterbrechungen aufwies²⁾ oder daß die Ziffern in Größe und Gestalt nicht den Formvorschriften genügten. Es war den einzelnen Statistischen Landesämtern freigestellt, Fehlerprotokolle anzulegen und Bereinigungen vorzunehmen. Immerhin erwiesen sich aus einer Gesamtheit von 2,9 Mill. Belegen, die der Verfasser beispielhaft auswertete, 46 200 (1,59 %) als korrekturbedürftig. Wer auf die Korrektur verzichtete, nahm außer den Informationsverlusten aus dem Stadium der Datensammlung noch diejenigen aus dem Leseprozeß in Kauf. Gegen einen Eingriff lassen sich gewiß dieselben Argumente wie oben vorbringen; der entscheidende Unterschied besteht aber darin, daß hier eine Originalinformation vorhanden war, deren Code lediglich nicht identifiziert werden konnte. Die

-
- 1) Auf Grund von Ergebnissen der Probeerhebung wurden in Kanada im Maschinenprogramm DIARY I Kontrollgrenzen für die gemeinsame Fehlerrate aus Nichtbeantwortung und Mängeln der Datenerfassung festgelegt. Die Werte betrugen z.B. für die Merkmale Geschlecht 5 %, Religion 10 %, Geschäftszweig und Beruf je 15 %, Selbständige (mit Ausnahme in der Landwirtschaft) 20 %. Korrekturen und erneute Maschinenläufe waren nur bei technischen Mängeln vorgesehen; bei Informationsverlusten infolge fehlender Angaben erfolgte kein Eingriff. Wenn die vorgegebenen Kontrollgrenzen wegen dieser Fehlerart ständig überschritten wurden, erfuhren sie eine Ausweitung: DIARY I, a.a.O., S.4-6 und 12.
 - 2) Sprick hatte die Idee, die vielfältigen Ausführungen handgeschriebener Zeichen nach dem Prinzip der Konturenverfolgung maschinell lesbar zu machen; Olf, F. K., Maschinen lesen Handschriften, in: IBM Nachrichten, 17. Jg., H. 181, S.422-428 und H. 182, S.486-493, Verlag IBM Deutschland, Sindelfingen 1967. Ders., Lesende Computer - Alternative der Datenerfassung, Datenverarbeitungs-Praxis, VRM-Schriftreihe, Köln 1973.

Korrektur der aus dem Leseprozeß ausgesteuerten Belege hätte einen Genauigkeitserfolg gebracht bei vertretbarem Arbeitsaufwand.

Handschrift-Lesefehler, die durch das Leseprogramm identifiziert werden können, sind korrigierbar. Diese Möglichkeit besteht nur sehr selten bei Substitutionen. Das sind Fehler, bei denen eine handschriftliche Ziffer - ohne daß eine Rückweisung durch das Gerät erfolgt - in eine andere Ziffer umgewandelt wird. Das ist offenbar darauf zurückzuführen, daß in der Logik des Lesers eine größere Ähnlichkeit zu einer anderen Musterziffer als der vorgesehenen erkannt wird.

Der Verfasser hat zwei Untersuchungen durchgeführt, von denen die erste die Bedeutung dieser Fehlerart im Zusammenhang mit anderen zeigt, während die zweite den Versuch macht, die Richtung zu bestimmen, in der die Substitutionen wirken.

Bei der Volkszählung 1970 war der Bestand an Personen im Alter von "100 und mehr Jahren" wesentlich überhöht. Insgesamt waren 579 Personen am äußersten oberen Rand der Altersverteilung aufbereitet worden.

Kontrolle der Personen im Alter von 100
und mehr Jahren

Fehlerart	Anzahl	%
Lesefehler	126	21,7
Ziffer(n) überlesen	68	11,7
Substitutionen	58	10,0
Signierfehler	272	47,0
Antwortfehler	101	17,4
kombinierte Fehler	5	0,9
ungeklärt	12	2,1
Personen im Alter von 100 und mehr Jahren	63	10,9
Insgesamt	579	100,0

Nur 63 Personen (10,9 %) erwiesen sich als "echte" Hundertjährige.

Beinahe die Hälfte der kontrollierten Fälle (47 %) bestand aus Signierfehlern, also fehlerhaften Übertragungen des Geburtsdatums in die Lesefelder. Der Antwortfehler wirkte sich zu 17,4 % aus. Er war in den meisten Fällen gleichzeitig ein "passiver" Signierfehler, da unlogische Antworten unbereinigt in die Lesefelder übernommen wurden. Hierbei spielte die Anweisung an die Signierer eine Rolle, bei den 1899 und früher Geborenen die Geburtsmonate 01 bis 12 um 20 (d.h. von 21 bis 32) zu erhöhen. Diese Vorschrift war sogar als deterministische Korrektur programmiert worden für Geburtsdaten nach dem 27. Mai 1970 (Zählungstichtag), die definitionsgemäß bei der Zählung nicht vorkommen durften. So entstanden "per Programm" Personen im Alter von "100 und mehr Jahren" einmal dadurch, daß einige Zähler Kleinkinder mit Geburtsdaten nach dem Zählungstichtag erfaßten oder übersehen wurde, daß eine Auskunftsperson anstelle ihres Geburtsdatums das Tagesdatum (nach dem Zählungstichtag) eingetragen hatte.

Lesefehler traten in 21,7 % der Fälle auf. Bei der Nichtidentifikation zu klein geschriebener Ziffern (11,7 %) stellte sich eine technische Eigenart des Beleglesers als besonders nachteilig heraus, die nächst links stehende Ziffer (hier des Geburtstages) automatisch nach rechts zu rücken (hier in die Zehnerstelle des Geburtsmonats), so daß - im Falle von 2 und 3 - bei bestimmten Geburtsjahren Personen dieses Alters "konstruiert" wurden. Substitutionen in 2 bzw. 3 in der Zehnerstelle des Geburtsmonats waren in 10 % der kontrollierten Fälle vertreten. Der Anteil kombinierter Fehler war sehr gering. Die Analyse zeigt eine Vielfalt an Fehlerquellen. Besonders interessant erscheint dem Verfasser, daß eine automatische Korrektur zu Verfälschungen führte, weil es nicht gelungen war, im technisch-organisatorischen Ablauf die zeitliche Abgrenzung der statistischen Gesamtheit durch den Zählungstichtag zu gewährleisten. In diesem Fall erfolgte nur noch die Bereinigung der Aggregate, nicht mehr der einzelnen Datensätze. Wie schon ausgeführt,¹⁾ gibt es auch kein Kriterium für die nachträgliche Umsetzung automatischer Aggregat-Bereinigungen in die einzelnen Datensätze. Um die vielseitige Verwendbarkeit der statistischen Ergebnisse von Volkszählungen nicht einzuschränken, ist hier keine andere Lösung möglich als die

1) Vgl. S.90/91.

Korrektur der einzelnen Datensätze auf Grund der Fehlerforschung und die Wiederholung aller Maschinenläufe bis zur Tabellierung. Der Aufwand dürfte im allgemeinen nicht vertretbar sein. Diese Erkenntnis führt den Verfasser zu einer stärkeren Differenzierung bei der Abgrenzung zwischen operativen und deskriptiven Verfahren. Wenn neue Informationen über Fehlerarten und Fehlerquellen in den verschiedenen Stufen des Aufbereitungsprozesses auftreten und organisatorisch noch die Möglichkeit sowie der Wille der Verantwortlichen zur Revision der jeweiligen "Ergebnisse" besteht, sind die Verfahren operativ. Wenn die "Ergebnisse" einer Stufe unverrückbar feststehen, kann nur noch konstatiert werden. In dem genannten Beispiel war das Bereinigungsverfahren für Landes- und Kreisergebnisse operativ, für alle übrigen Ergebnisse dagegen bereits deskriptiv.

Mit Hilfe der Redundanzen im Geburtsmonat ließ sich ein Teil der Substitutionsrate nachweisen. Wie oben ausgeführt, waren im Geburtsmonat die Ziffern 01 bis 12 sowie 21 bis 32 zulässig. Durch Substitution entstandene unzulässige Ziffern konnten automatisch erfolgreich kontrolliert werden. Auf 63 400 Belegen¹⁾ einer Gesamtmasse von 70 382 war der Geburtsmonat nachgewiesen. Von den rd. 127 000 Ziffern waren 74 (0,06 %) Substitutionen. Da es von der automatischen Kontrolle nicht erkennbare Substitutionen in gültige Codes gab, mußte die effektive Rate höher liegen. Eine sehr viel aufwendigere Methode besteht darin, die maschinell erkannten Ziffern zu protokollieren und einen manuellen Vergleich mit den handschriftlichen Ziffern durchzuführen. Der Verfasser hat auf 3 704 Belegen je 4 Ziffern (Geburtsmonat und Geschäftszweig) kontrolliert. Unter den 14 816 Ziffern befanden sich 20 (0,13 %) Substitutionen. Davon verfälschten etwa die Hälfte das Ergebnis. Da die Belege von unterschiedlichen Arbeitskräften signiert wurden, scheint die Gefahr einer systematischen Verzerrung nicht gegeben zu sein, wenngleich die Zahl der Fälle natürlich gering ist.

Wenn die Substitutionen in bestimmte Richtungen tendieren, liegt der Gedanke an operative Korrekturen nahe. Die weiter oben genannten 74 Substitutionen verteilen sich wie folgt:

1) Vgl. Anhang S.180.

Substitutionen im Geburtsmonat

Ziffer		Anzahl
signiert	gelesen	
0	1	20
0	3	2
0	4	3
0	5	1
0	6	2
0	8	27
0	9	4
1	2	1
1	6	2
2	3	2
2	5	1
2	6	1
2	7	5
2	9	1
6	0	2
Insgesamt		74

In 27 Fällen (36,5 %) wurde 0 zu 8, in 20 Fällen (27 %) 0 zu 1. Die 0 erwies sich damit als relativ fehleranfällige Ziffer. Die operativen Eingriffe beziehen sich in diesem Fall auf die Logik des Lesegeräts, für die Fehlerbereinigung bei statistischen Erhebungen sind sie irrelevant, weil der Fehler nicht zu lokalisieren ist. Zwei technische Eingriffe sind möglich, um nicht hinnehmbare Substitutionsraten zu verkleinern. Die Ziffern können mehrfach gelesen werden, wodurch allerdings die Leseleistung des Geräts beeinträchtigt wird. Die durchschnittliche Stundenleistung hatte in Nordrhein-Westfalen 7 150 Belege betragen. Außerdem kann die Toleranz des Lesers beim Vergleich der handschriftlichen Ziffern mit den Musterziffern eingeschränkt werden. Dadurch wird aber ein höherer Anteil an Ziffern als nicht identifizierbar protokolliert. Das Statistische Bundesamt sah von einer Veränderung des Leseprogramms ab, da es die Substitutionsrate für kleiner als den Lochfehler bei früheren Zählungen hielt. Um die Auswirkungen der Beleglesung auf die Ergebnisse von Volkszählungen verlässlicher abzuschätzen, bedarf es sicherlich umfangreicher Forschungsarbeiten. Das Beispiel der Personen im Alter "über 100 Jahre" zeigt aber, daß der Lesefehler im Rahmen der übrigen Fehlerarten ernst zu nehmen ist.

9 Schlußbetrachtung

Der Konsument sollte die Qualität der Statistiken aus Volkszählungen nicht überschätzen. Die Perfektion der Ergebnistabellen täuscht. Sie bietet dem Konsumenten Annehmlichkeiten und beweist dem Produzenten konsistente Aufbereitung. Die Statistiken sollten nicht ohne Genauigkeitsmessungen oder wenigstens -beschreibungen veröffentlicht werden. Das in der Bundesrepublik am dringlichsten zu lösende Problem erscheint dem Verfasser die stichprobenhafte Erhebung einer Reihe von Merkmalen zu sein. Den Verwendern regional und sachlich sehr tief gegliederter Ergebnisse sollte möglichst schnell nachgewiesen werden, daß die systematischen Erfassungs-, Antwort-, Signier-, Loch- bzw. Lesefehler und Verfälschungen durch die automatische Fehlerbereinigung ein unteres Qualitätslimit setzen, das durch den Gesamtfehler von Stichproben im Umfang von 10 bis 20 % kaum weiter unterschritten wird, vorausgesetzt, die Erhebungsphase kann besser kontrolliert werden und Verfahren Statistischer Qualitätskontrollen werden im Signierprozeß eingesetzt, so daß im Vorfeld der automatischen Datenverarbeitung eine Grenzqualität (maximaler Durchschlupf) gewährleistet ist.

Der Verfasser schätzt trotz vieler Vorkehrungen seitens der amtlichen Statistik den qualitätsverbessernden Effekt der Gemeinden als Erhebungsstellen nicht hoch ein. Oft wahren sie nicht einmal ihr dringlichstes Interesse, für den Finanzausgleich auf höchste Erfassungsgenauigkeit zu achten. Langfristig muß das Vertrauen der Bevölkerung in die Geheimhaltung der Daten zumindest erhalten werden, damit die Erfassungs- und Antwortgenauigkeit nicht weiter abnimmt. Öffentlichkeitsarbeit ist in diesem Zusammenhang unabdingbar. Der repräsentative Teil der Zählung 1970 enthielt einige "schwere" Fragen. Antwortausfälle oder Inkonsistenzen bei diesen Fragen belasteten die Fehlerbereinigung am stärksten. Besonders hier waren Serieneinsetzungen aus den Decks zu beobachten. Hier wäre vermutlich ein höherer Auswahlatz als 10 % von Vorteil gewesen. Wenn an grundlegende Veränderungen des Zählungskonzepts aber nicht gedacht ist, wäre im Rahmen der Korrekturen eine Speicherung mehrerer Informationen je Deckzelle vorzusehen. Unter Genauigkeits-

aspekten erscheint eine Unterbrechung des automatischen Ablaufs zum Zwecke manueller Überkorrekturen unverzichtbar.

Der Verfasser ist äußerst skeptisch gegenüber der Anwendung von Verteilungen aus früheren Volkszählungen für die Fehlerbereinigung des aktuellen Projekts. Nicht nur, daß ein beträchtlicher Teil der historischen Verteilung selbst durch Bereinigung zustande gekommen sein kann, es erscheint durchaus möglich, daß bei relativ "statischen" Merkmalen die minimalen Veränderungen im Zeitablauf "herauskorrigiert" werden.

Da viele in der Untersuchung analysierte Verfahren quantitative Merkmale voraussetzen, sind sie für Volkszählungen nicht brauchbar. Es besteht kein Zweifel, daß die Bedeutung des systematischen Fehlers bei vorwiegend qualitativen Merkmalen von Volkszählungen nur entscheidend gemildert werden kann durch Überwachung der organisatorischen Konzepte. Hier nützt die Stichprobentheorie wenig; sie wird benötigt, wenn Kontrollerhebungen und Schattenaufbereitungen für die Genauigkeitsmessung einzelner Maßnahmen zur Qualitätsverbesserung ausgewertet werden. Solche Untersuchungen fehlen in der Bundesrepublik weitgehend.

Der Verfasser hält es für richtig, daß die Vorschläge von Fellegi/Holt zur Fehlerbereinigung, die auf dem Konzept der Kontrollen als Wertemengen beruhen, geprüft werden. Ohne empirische Kenntnisse über den maschinellen Suchaufwand der von den Autoren entwickelten sequentiellen Korrekturverfahren, ist die Methode nicht zu bewerten. Allerdings dürfte auch bereits die theoretisch geforderte Vollständigkeit des Kontrollrahmens schwer zu gewährleisten sein. Wenn dieses System, wie vom Verfasser vermutet, auf Volkszählungen nicht anwendbar ist, muß weiterhin an der Schichtenbildung (besonders im Falle fehlender oder falscher Merkmale) und an der Fehlerallokation gearbeitet werden. Neben Genauigkeitsverbesserungen ist hier eine Reduktion der Maschinenzeit durch Verkleinerung der Zahl der Iterationen möglich.

Schließlich sollten noch ein paar Schlaglichter auf den organisatorischen Hintergrund der Datenverarbeitung bei Statistiken in der Bundesrepublik geworfen werden. Innerhalb der Statistischen Landesämter besteht ein "Programmierverbund", weil keine Behörde über ausreichende Programmierkapazität verfügt, um alle Programme eines Projektes zu bearbeiten. Der Verfasser hält die Kommunikationsprobleme dieses Verbundes für so schwerwiegend, daß sie die Leistungsfähigkeit der amtlichen Statistik entscheidend mindern.

Für den Erfolg der statistischen Arbeit ist Kooperation zwischen Fachstatistikern und Organisatoren bzw. Systemanalytikern der maschinellen Abteilungen wichtig. Die zentrale Bedeutung des Computers für den Zählungsablauf setzt aber auch profunde Kenntnisse der EDV beim Fachstatistiker voraus. Er kann sonst kaum verhindern, daß die maschinellen Abteilungen die Konzepte der automatischen Fehlerbereinigung bestimmen. Der Fachstatistiker sollte immer in der Lage sein zu erkennen, wann technisch einfachere Problemlösungen qualitativ besseren vorgezogen werden sollen, nur weil diese etwa leichter zu programmieren sind. Von großem Nutzen für die Fachstatistik ist eine vollständige Dokumentation von Programmablaufplänen, Datenflußplänen, Testergebnissen usw. Ohne diese Unterlagen sind Störungen des Arbeitsablaufs nur schwer zu beheben, die unweigerlich zu Aktualitätsverlusten führen.

Außer von der wissenschaftlichen Grundlagenforschung hängt die Aktualität und die Qualität von Zählungsergebnissen heute wesentlich davon ab, wie schnell und präzise die technischen und methodischen Probleme der Zählung analysiert, definiert und in fehlerfreie Maschinenprogramme umgesetzt werden können. Der Optimismus erscheint zulässig, daß beide Ziele bei Volkszählungen in der Bundesrepublik in Zukunft noch besser zu erreichen sind.

Literaturverzeichnis1. Bücher und Zeitschriften

- Akers, D. S.,
Larmon, E. A.,
Indians and Smudges on the Census Schedule, in: Proceedings of the Social Statistics Section, Hrsg. American Statistical Association, Washington, D.C. 1967, S.369-373.
- Anderson, O.,
Probleme der statistischen Methodenlehre in den Sozialwissenschaften, 3. wes. umgearb. u. erw. Aufl., Physica, Würzburg 1957.
- Bauknecht, K.,
Nef, W. (Hrsg.),
Digitale Simulation, in: Lecture Notes in Operations Research and Mathematical Systems, H. 51, Springer, Berlin-Heidelberg-New York 1971.
- Becker, J.,
Woit, J.,
Zajonc, H.,
Der Einsatz der Elektronischen Datenverarbeitung in der Bundesrepublik Deutschland 1970, Hrsg. Ausschluß für wirtschaftliche Verwaltung eV (AWV) und Gesellschaft für Kernforschung mbH, Frankfurt-Karlsruhe 1971.
- Bose, R. C.,
Error Detecting and Error Correcting Indexing Systems for Large Serial Numbers, in: Review of the International Statistical Institute, Vol. XXXIV, Book 3, The Hague 1966, S.334-340.
- Bose, R. C.,
Error Detecting and Error Correcting Indexing Systems for Large Serial Numbers (Abstract), in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.472.
- Brennan, J. F.,
Computer "reparieren" sich selbst - Entwicklung von Verfahren zur automatischen Fehlersuche und -korrektur, in: IBM Nachrichten, 20. Jg., H. 199, Verlag IBM Deutschland, Sindelfingen 1970, S.7-13.
- Buck, S. F.,
A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer, in: Journal of the Royal Statistical Society, Ser. B, Vol. 22, London 1960, S.302-306.

- Bugmann, A., Statistische Techniken und Datenverarbeitung bei Großzählungen, in: Schweizerische Zeitschrift für Volkswirtschaft und Statistik, Hrsg. Schweizerische Gesellschaft für Statistik und Volkswirtschaft, 107. Jg., H. 3, Bern 1971, S.621-626.
- Cassel, P. G.,
Fisz, A., Statistical Problems in Automatic Control of Data, in: Statistisk Tidskrift, Hrsg. Statistiska Centralbyrån, H. 2, Stockholm 1970, S.96-99.
- Dalenius, T., Automatic Estimation of Missing Values in Censuses and Sample Surveys, in: Statistisk Tidskrift, Hrsg. Statistiska Centralbyrån, H. 7, Stockholm 1962, S.395-400.
- Deininger, R., Maschinelle Beleglesung bei der Volkszählung 1970 in der Bundesrepublik Deutschland, in: IBM Nachrichten, 22. Jg., H. 210, Verlag IBM Deutschland, Sindelfingen 1972, S.121-126.
- Deininger, R.,
Kerler, E., Maschinelle Datenprüfung in der amtlichen Statistik - Automatische Korrekturverfahren bei der Gebäude- und Wohnungszählung 1968, in: Allgemeines Statistisches Archiv, Bd 55, H. 3, Göttingen 1971, S.290-303.
- Deming, W. E., Some Theory of Sampling, Wiley & Sons, Chapman & Hall, New York/London 1950.
- Dixon, W. J., Analysis of Extreme Values, in: The Annals of Mathematical Statistics, Vol. 21, Waverly Press, Baltimore 1950, S.488-506.
- Dixon, W. J., Ratios Involving Extreme Values, in: The Annals of Mathematical Statistics, Vol. 22, Waverly Press, Baltimore 1951, S.68-78.
- Dumas, L., Auswertungsmöglichkeiten des Zählungswerkes 1968/70 - Überlegungen unter dem Aspekt der Stadtentwicklung -, in: Hamburg in Zahlen, Hrsg. Statistisches Landesamt Hamburg, Jg. 1972, März-Heft, S.75-88.
- Fellegi, I. P., The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences, in: International Statistical Review, Vol. 41, No. 1, Edinburgh 1973, S.1-14.
- Fellegi, I. P.,
Goldberg, S. A., Some Aspects of the Impact of the Computer on Official Statistics, in: Bulletin of the International Statistical Institute, Proceedings of the 37th Session, Vol. XLIII, Book 1, London 1969, S.157-177.

- Fellegi, I. P.,
Holt, D.,
A Systematic Approach to Automatic Editing and Imputation, in: Bulletin of the International Statistical Institute, Proceedings of the 39th Session, Vol. XLIV, Book 1, Wien 1973, S.364-376.
- Flaskämper, P.,
Allgemeine Statistik, bericht. Nachdruck der 2. erg. Aufl. 1949, Meiner, Hamburg 1962.
- Freund, R. J.,
Hartley, H. O.,
A Procedure for Automatic Data Editing, in: Journal of the American Statistical Association, Vol. 62, No. 319, Menasha, Wisc. 1967, S.341-352.
- Fürst, G.,
Maschinenverwendung und Automatisierung in der Statistik, in: Allgemeines Statistisches Archiv, Bd 43, H. 4, München 1959, S.313-315.
- Gumbel, E. J.,
Statistics of Extremes, Columbia University, Fourth Printing, New York 1967.
- Hansen, M. H.,
Hurwitz, W. N.,
Bershad, M. A.,
Measurement Errors in Censuses and Surveys, in: Bulletin of the International Statistical Institute, Proceedings of the 32nd Session, Vol. XXXVIII, Book 2, Tokyo 1961, S.359-374.
- Hansen, M. H.,
Hurwitz, W. N.,
Madow, W. G.,
Sample Survey Methods and Theory, Vol. I, Methods and Applications, Vol. II, Theory, Wiley & Sons, New York 1953.
- Hansen, M. H.,
Hurwitz, W. N.,
Marks, E. S.,
Maulding, W. P.,
Response Errors in Surveys, in: Journal of the American Statistical Association, Vol. 46, No. 254, Menasha, Wisc. 1951, S.147-190.
- Hansen, M. H.,
Waksberg, J.,
Research on Non-Sampling Errors in Censuses and Surveys, in: Review of the International Statistical Institute, Vol. 38, No. 3, The Hague 1970, S.317-332.
- Hartley, H. O.,
Hocking, R. R.,
The Analysis of Incomplete Data, in: Biometrics, Vol. 27, Hrsg. Biometric Society, Raleigh, NC. 1971, S.783-823.
- Hawkins, D. M.,
The Detection of Errors in Multivariate Data Using Principal Components, in: Journal of the American Statistical Association, Vol. 69, No. 346, Menasha, Wisc. 1974, S.340-344.
- Herberger, L.,
Verfahrenskontrollen zur Prüfung der Vollständigkeit und Genauigkeit bei Volks- und Berufszählungen, in: Allgemeines Statistisches Archiv, Bd 54, H. 1, Göttingen 1970, S.55-75.

- Hocking, R. R.,
Huddleston, H. F.,
Hunt, H. H., A Procedure for Editing Survey Data,
in: Journal of the Royal Statistical
Society, Applied Statistics, Ser. C,
Vol. 23, No. 2, London 1974,
S.121-133.
- Hofsten, E. v., Population Registers and Computers.
New Possibilities for the Production
of Demographic Data, in: Review of
the International Statistical
Institute, Vol. XXXIV, Book 2,
The Hague 1966, S.186-194.
- Kendall, M. G.,
Buckland, W. R., A Dictionary of Statistical Terms,
3. Aufl., Oliver & Boyd, Edinburgh 1972.
- Kivivuori, R., A Method for Checking Numerical Codes
Using the 1401, in: Nordisk Tidskrift
for informasjonsbehandling, Bd 1,
H. 1, Lund 1961, S.48-53.
- Klaus, G. (Hrsg.), Wörterbuch der Kybernetik, Dietz,
Berlin 1968.
- Koelle, H. H., Die Anwendung der Simulation als Ent-
scheidungshilfe, in: IBM Nachrichten,
21. Jg., H. 208, Verlag IBM
Deutschland, Sindelfingen 1971,
S.873-879.
- Manescu, M., Electronic Data Processing of the
Population and Housing Census in
Romania, in: Economic Computation
and Economic Cybernetics, Studies
and Research, Hrsg. The Centre of
Economic Computation and Economic
Cybernetics, H. 4, Bucharest 1968,
S.29-54.
- Minton, G., Inspection and Correction Error in
Data Processing, in: Journal of the
American Statistical Association,
Vol. 64, No. 328, Menasha, Wisc. 1969,
S.1256-1275.
- Morgenstern, O., Über die Genauigkeit wirtschaftlicher
Beobachtungen, 2. erw. Aufl.
Physica, Wien-Würzburg 1965.
- Naus, J. I., Data Quality Control and Editing,
Statistics: Textbooks and Monographs,
Vol. 10, Marcel Dekker, New York 1975.
- Naus, J. I.,
Johnson, T. G.,
Montalvo, R., A Probabilistic Model for Identifying
Errors in Data Editing, in: Journal
of the American Statistical Association,
Vol. 67, No. 340, Menasha, Wisc. 1972,
S.943-950.
- Nordbotten, S., Measuring the Error of Editing the
Questionnaires in a Census, in:
Journal of the American Statistical
Association, Vol. 50, No. 270,
Menasha, Wisc. 1955, S.364-369.

- Nordbotten, S., On Errors and Optimal Allocation in a Census, in: Skandinavisk Aktuarietidskrift, 40. Jg., H. 1, Uppsala 1957, S.1-10.
- Nordbotten, S., Automatic Editing of Individual Statistical Observations, Statistical Standards and Studies, No. 2, Hrsg. United Nations, Conference of European Statisticians, New York 1963.
- Nordbotten, S., The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means for Improving the Quality of Statistics, in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.417-441.
- Olf, F. K., Maschinen lesen Handschriften, in: IBM Nachrichten, 17. Jg., H. 181, S.422-428 und H. 182, S.486-493, Verlag IBM Deutschland, Sindelfingen 1967.
- Olf, F. K., Lesende Computer - Alternative der Datenerfassung, Datenverarbeitungs-Praxis, VRM-Schriftenreihe, Köln 1973.
- O'Reagan, R. T., Possible Techniques for Computer Editing of Data, US Bureau of the Census, Washington, D.C 1965.
- Pfanzagl, J., Artikel "Fehlertheorie" (S.482), in: Handwörterbuch der Sozialwissenschaften (HdSW), Bd 3, Fischer, Stuttgart, Mohr, Tübingen, Vandenhoeck & Ruprecht, Göttingen 1961.
- Pritzker, L.,
Ogus, J.,
Hansen, M. H., Computer Editing Methods. Some Applications and Results, in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.442-466.
- Putz, F., Der Beitrag der Zähler zur operativen Verfahrenskontrolle bei einer Großzählung, in: Allgemeines Statistisches Archiv, Bd 54, H. 3, Göttingen 1970, S.295-306.
- Sauter, U., Die moderne statistische Qualitätskontrolle unter besonderer Berücksichtigung ihrer Bedeutung für die Wirtschaftlichkeit der Unternehmung, Diss., Mannheim 1959.

- Schubnell, H.,
Herberger, L. u.a., Die Vorbereitung des Zählungs-
werkes 1970, Teil 2, Die Volks- und
Berufszählung, in: Wirtschaft und
Statistik, H. 10, Wiesbaden 1967,
S.549-559.
- Schubnell, H.,
Herberger, L., Die Volkszählung am 27. Mai 1970,
in: Wirtschaft und Statistik, Hrsg.
Statistischen Bundesamt, H. 4,
Stuttgart-Mainz 1970, S.179-185.
- Schuff, H. K., Fehler in informationsverarbeitenden
Systemen und ihre Bedeutung, Teil II,
in: Elektronische Datenverarbeitung,
H. 2, Braunschweig 1964, S.67-75.
- Statistisches Bundesamt, Wiesbaden (Hrsg.),
Stichproben in der amtlichen Statistik,
Kohlhammer, Stuttgart-Mainz 1960.
- Organisation und Technik des Volks-
zählungswerkes 1950, Reihe:
Statistik der Bundesrepublik
Deutschland, Bd 31, Kohlhammer,
Stuttgart-Köln 1956.
- Volks- und Berufszählung vom
6. Juni 1961, H. 1, Die methodischen
Grundlagen der Volks- und Berufs-
zählung 1961, Reihe: Bevölkerung und
Kultur (Fachserie A), Kohlhammer,
Stuttgart-Mainz 1967.
- Organisation und Technik des Zählungs-
werkes 1960/62, Kohlhammer,
Stuttgart-Mainz 1968.
- Das Arbeitsgebiet der Bundesstatistik,
Kohlhammer, Stuttgart-Mainz 1971.
- Storch, H., Die Religionszugehörigkeit der Be-
völkerung Nordrhein-Westfalens am
6. Juni 1961, in: Statistische Rund-
schau für das Land Nordrhein-West-
falen, Hrsg. Statistisches Landes-
amt Nordrhein-Westfalen, 15. Jg.,
H. 10, Düsseldorf 1963, S.217-223.
- Storch, H., Zusammenfassender Bericht über die
Erfahrungen bei der Volks- und Be-
rufszählung 1961, Hrsg. Statistisches
Landesamt Nordrhein-Westfalen,
Düsseldorf 1965.
- Swoboda, H., Knaurs Buch der modernen Statistik,
Droemer Knaur, München-Zürich 1971,
- Szameitat, K., Möglichkeiten und Grenzen der Automa-
tisierung in der Statistik, in: All-
gemeines Statistisches Archiv, Bd 43,
H. 4, München 1959, S.316-333.

- Szameitat, K.,
Deininger, R., Some Remarks on the Problem of Errors in Statistical Results, in: Bulletin of the International Statistical Institute, Proceedings of the 36th Session, Vol. XLII, Book 1, Sydney 1967, S.66-89.
- Szameitat, K.,
Deininger, R., Die Vorbereitung des Zählungswerkes 1970, Teil 1, Das Gesamtkonzept, in: Wirtschaft und Statistik, H. 8, Wiesbaden 1967, S.443-453.
- Szameitat, K.,
Zindler, H. J., The Reduction of Errors in Statistics by Automatic Corrections, in: Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Vol. XLI, Book 1, Beograd 1966, S.395-417.
- Terry, M. E., The Principles of Statistical Analysis Using Large Electronic Computers, in: Bulletin of the International Statistical Institute, Proceedings of the 34th Session, Vol. XL, Book 1, Ottawa 1963, S.547-552.
- Thorelli, L. E., Automatic Correction of Errors in Text, in: Nordisk Tidskrift for informasjonsbehandling, Bd 2, H. 1, Lund 1962, S.45-52.
- Tukey, J. W., The Future of Data Analysis, in: The Annals of Mathematical Statistics, Vol. 33, Waverly Press, Baltimore 1962, S.1-67.
- Uhlmann, W., Statistische Qualitätskontrolle, Teubner, Stuttgart 1966.
- Ungern-Sternberg, R.v.
und Schubnell, H., Grundriß der Bevölkerungswissenschaft (Demographie), Piscator, Stuttgart 1950.
- US Bureau of the Census, Washington D.C. (Hrsg.),
United States Censuses of Population and Housing 1960, Processing the Data, Washington, D.C. 1962.
Evaluation and Research Program of the US Censuses of Population and Housing 1960, Accuracy of Data on Population Characteristics as Measured by Reinterviews, Washington, D.C. 1964.
Evaluation and Research Program of the US Censuses of Population and Housing 1960, Accuracy of Data on Population Characteristics as Measured by CPS-Census Match, Washington, D.C. 1964.

US Bureau of the Census (Fortsetzung)

United States Censuses of Population and Housing 1960, Quality Control of Preparatory Operations, Microfilming, and Coding, Washington, D.C. 1965.

United States Censuses of Population and Housing 1960, Quality Control of the Field Enumeration, Washington, D.C. 1967.

United States Censuses of Population and Housing 1970, Procedural History PHC(R)-1G, Advance Issuance of Chapter 15 (Population and Housing Items on the US Household Questionnaires), Washington, D.C. 1974.

United States Censuses of Population and Housing 1970, Procedural History PHC(R)-1H, Advance Issuance of Chapters 7 and 8 (The Coverage Improvement Program, Processing the Data), Washington, D.C. 1975.

Waksberg, J.,
Pritzker, L.,

Changes in Census Methods, in: Journal of the American Statistical Association, Vol. 64, No. 328, Menasha, Wisc. 1969, S.1141-1149.

Wallraf, R.,

Die Sünden der Statistiker, in: Capital, Das deutsche Wirtschaftsmagazin, H. 12, Hamburg 1969, S.100-106.

Weichselberger, K.,

Kontrollen der Ergebnisse von Volkszählungen, Habil.-Schrift, Köln 1962.

Weichselberger, K.,

Genauigkeitsansprüche, Fehler und Kontrollen bei Volkszählungen, in: Allgemeines Statistisches Archiv, Bd 54, H. 1, Göttingen 1970, S.1-28.

Zindler, H. J.,

Die Auswirkungen maschineller Korrekturverfahren, in: Metrika, Internationale Zeitschrift für theoretische und angewandte Statistik, Vol. 11, Wien-Würzburg 1967, S.206-229.

2. Sonstige Quellen

- Deininger, R.,
Kerler, E., Die Praxis automatischer Korrekturen und ihre Anwendung bei der Gebäude- und Wohnungszählung 1968, hektographierte Arbeitsunterlage, Hrsg. Statistisches Landesamt Baden-Württemberg, Stuttgart 1969.
- Fellegi, I. P., Automatic Editing and Imputation of Quantitative Data, hekt. Arbeitsexemplar, Toronto 1975.
- Fellegi, I. P., Automatic Editing and Imputation of Quantitative Data (Summary), Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975.
- Nargundkar, M. S.,
Joshi, G. B., Non - Response in Sample Surveys, Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975.
- Palmer, S.,
Jones, Ch., A Look at Alternate Imputation Procedures for CPS Noninterviews, hekt. Veröffentlichung, Hrsg. US Bureau of the Census (Statistical Methods Division), Washington, D.C. 1967.
- Sanchez-Crespo, J. L., Notes on the Accuracy, Precision and Reliability of Statistical Data, Diskussionsunterlage der 40. Sitzung des Internationalen Statistischen Instituts, Warschau 1975.
- Dominion Bureau of Statistics, Final Report of the DIARY I Task Force: Detailed Specifications, hekt. Arbeitsunterlage, Ottawa 1970.
- Dominion Bureau of Statistics, Report of the DIARY II Task Force: Specifications for the DIARY II Programme, hekt. Arbeitsunterlage, Ottawa 1971.
- Österreichisches Statistisches Zentralamt, Wien, Bericht über das Expertentreffen zur Beratung von Volkszählungsfragen vom 4. bis 6. Nov. 1968 in Wien, Protokoll veröffentlicht am 19. Nov. 1968.

Statistisches Bundesamt, Wiesbaden,

Schreiben vom 10.10.1969, VIII V 2/00,
Br/Gr: Manuelle und maschinelle Aus-
wertung der 6. Probeerhebung zur
Volkszählung 1970 mit Anlagen und
Tabellen.

Vermerk vom 12.11.1969, VII A 21,
Kp/Hb: Feststellung der Wohnbevölke-
rung bei der Volkszählung 1970.

Schreiben vom 19.2.1970, VIII V 4/00,
Schk/Br: Verfahren bei Mehrfach-Mar-
kierungen.

Schreiben vom 19.4.1971, VIII V 4/00,
Li/L/Schä: Zuordnung zur Wohnbevölke-
rung.

Schreiben vom 17.9.1971, VIII V 4/00,
Stg/Kg: Manuelle Überkorrekturen.

United Nations, Principles and Recommendations for
National Population Censuses, in:
Statistical Papers, Ser. M, No. 27,
New York 1958.

United Nations, Economic and Social Council, Statistical
Commission and Economic Commission for
Europe, Conference of European
Statisticians, 22nd Plenary Session,
Genf 1974, Accuracy and Statistics,
Arbeitspapier, bearb. von Statistics Canada.

US Bureau of the Census, 1970 Census of Population, Job
Content Computer Edits and Allocation
Specifications, hekt. Arbeitsunterlage,
Washington, D.C. 1970.

Gesetz über die Statistik für Bundeszwecke (StatGes) vom
3. September 1953 (BGBl. I S.1314).

Gesetz über eine Volks-, Berufs- und Arbeitsstättenzählung
(Volkszählungsgesetz 1970) vom
14. April 1969 (BGBl. I S.292).

Mündliche Anfrage des Abg. Strohmayer betr. Repräsentativ-
befragung in der Volkszählung 1970,
Bundestags-Drucksache VI/415 -
Schriftliche Antwort, 35. Sitzung,
S.1716 A.

Anhang

Schlüssel für die Volkszählung 1961 in der Bundesrepublik

1 0 2 2 7 4 2 9 2 2

Art der Statistik:	Totalzählung (1)
Charakterisierung der statistischen Unterlagen:	Keine Verwaltungsschriftstücke (o)
Erhebungsverfahren:	Erhebung durch Zähler (2)
Erhebungskontrollen:	Kontrolle des totalen und teilweisen Informationsausfalls (2)
Manuelle Belegkontrollen:	Vergleich mit anderen Zählungen und totale Signier- und Kombinations- kontrollen (7)
Manuelle Datenbehandlung:	Signieren aller verbalen Informa- tionen (4)
Kontrolle der manuellen Datenbehandlung:	Stichprobenkontrolle mit größerem Auswahlsatz (2)
Aufbereitungsverfahren:	EDV, Signierkontrollen (9)
Automatische Kontrollen:	Kombinationskontrollen mit manueller Überkorrektur (2)
Tabellenerstellung:	Nach der maschinellen Tabellierung manuelle Tabellenerstellung (2)

Quelle: Some Remarks on the Problem of Errors in Statistical Results, in: Bulletin of the International Statistical Institute, Proceedings of the 36th Session, Vol. XLII, Book 1, Sydney 1967, S.80 ff.

Merkmalskatalog der Volkszählung 1970
(Total- und Repräsentativteil)

1. Bevölkerungsstatistische Merkmale

- a) Geschlecht
- b) Geburtsjahr / Alter
- c) Familienstand
- d) Stellung innerhalb des Haushalts
- e) Religionszugehörigkeit
- f) Staatsangehörigkeit
- g) Zugehörigkeit zur Wohnbevölkerung
- h) Vertriebenen-, Flüchtlingseigenschaft

2. Merkmale der Geburtenstatistik

- a) Eheschließungsjahr und frühere Ehe
- b) Geburtsjahr aller lebendgeborenen ehelichen Kinder

3. Erwerbs- und wirtschaftsstatistische Merkmale

- a) Überwiegender Lebensunterhalt
- b) Beteiligung am Erwerbsleben / Arbeitssuche
- c) Name und Anschrift des Betriebes bzw. der
Schule / Hochschule
- d) Verkehrsmittel für den Weg zur Arbeits- bzw.
Ausbildungsstätte
- e) Zeitaufwand
- f) Wirtschaftszweig des Betriebes
- g) Stellung im Beruf
- h) Wochenarbeitszeit
- i) Weitere Tätigkeit in der Landwirtschaft
- k) Frühere Erwerbstätigkeit
- l) Jahr der Aufgabe der Erwerbstätigkeit
- m) Bezeichnung und Beschreibung der ausgeübten Tätigkeit
- n) Maschinenbedienung
- o) Höhe des Nettoeinkommens aus Erwerbstätigkeit
- p) Art der Tätigkeit (leitend oder aufsichtsführend)
- r) Anzahl der im Betrieb beschäftigten Personen
- s) darunter Lohn- oder Gehaltsempfänger

4. Bildungsstatistische Merkmale

- a) Schulabschluß
- b) Gegenwärtiger Schulbesuch
- c) Dauer der praktischen Berufsausbildung
- d) Art der praktischen Berufsausbildung
- e) Art der berufsbildenden Schule / Hochschule
- f) Hauptfachrichtung des höchsten Schulabschlusses
- g) Dauer der Berufsausbildung / Hochschulstudium
- h) Jahr des höchsten Schulabschlusses

5. Angaben über bewirtschaftete Kleinflächen

6. Telefonanschluß

Straße Hausnr.

Zahlbezirk

Eigener Telefonanschluß in der Wohnung (Nr.)

1 Geschlecht männlich
weiblich

2 Geburtsdatum Tag .. Monat .. Jahr ..

3 Familienstand ledig
verheiratet
verwitwet
geschieden

4 Stellung innerhalb des Haushalts Haushaltsvorstand (HV)
Ehegatte des HV
(Schwieger-) Sohn, Tochter des HV*)
Enkel, Urenkel des HV*)
Vater, Mutter des HV*)
Großvater, -mutter des HV*)
andere mit dem HV verwandte oder verschwägte Person
mit dem HV nicht verwandte Person

5 Religionszugehörigkeit Evang. Kirche (ohne Freikirche)
Evang. Freikirche
Röm.-kath. Kirche
sonstige christliche Gemeinsh.
Jüdische Rel.-Gemeinschaft
gemeinschaftslos
sonstige religiöse Gemeinschaften
Wenn sonstige welche?

6 Staatsangehörigkeit deutsch
nicht deutsch
Wenn nicht deutsch, welche?

7 Haben Sie noch eine andere Wohnung od. Unterkunft? nein
ja
a) Gehen Sie von der anderen Wohnung/Unterkunft aus zur Arbeit oder Schule/Hochschule? ja
nein
b) Leben Sie überwiegend in der anderen Wohnung/Unterkunft? ja
nein
c) Sind Sie in der anderen Wohnung/Unterkunft behördlich gemeldet? ja, mit Hauptwohn.
ja, mit Nebenwohn.
nicht gemeldet
d) Befindet sich die andere Wohnung/Unterkunft im Ausland? ja
nein

8 Wovon leben Sie überwiegend? Erwerbs-/Berufstätigkeit
Arbeitslosengeld/-hilfe
eigene Rente, Pension
Unterhalt durch Eltern, Ehemann usw.
eig. Vermögen, Vermietung, Zinsen, Altenteil
Sozialhilfe und sonstige Unterstützungen

9 Besuchen Sie gegenwärtig eine Schule? Volksschule
Berufsschule
Realschule
Gymnasium
Berufsfach-/Fachschule*)
Ingenieurschule
Hochschule
*) auch Technikerschule

10 Haben Sie einen Schulabschluß? Volksschule
Berufsschule
Mittlere Reife
Abitur
Berufsfach-/Fachschule*)
Ingenieurschule
Hochschule (auch Lehrerausb.)

11 Sind Sie erwerbstätig
Landwirt
mithelfend im Familienbetrieb
arbeitslos/arbeitsuchend
nicht erwerbstätig
Hausfrau
Berufs-/Zeitsoldat

Für Erwerbstätige (auch Selbständige, Landwirte, Mithelfende) sowie Schüler und Studierende

12 Wo arbeiten Sie/wo gehen Sie zur Schule/Hochschule?

Name des Betriebes bzw. der Schule/Hochschule

Gemeinde und Kreis der Arbeitsstätte, Filiale, Baustelle oder Schule

Straße Hausnummer

13 Welches Verkehrsmittel benutzen Sie hauptsächlich (längste Strecke) auf dem Hinweg zur Arbeit oder Schule/Hochschule?
kein Verkehrsmittel (zu Fuß)
Eisenbahn
Straßenbahn
S-Bahn, U-Bahn, Hochbahn
Werkbus, Schulbus
öffentl. oder priv. Linienbus
Motorrad, Moped
Fahrrad, sonst. Verkehrsmittel (ohne Auto)
Auto { Selbstfahrer
Mitfahrer
Bei Autobenutzung: { bis 10 km
Entfernung (Hinweg) { über 10 km

14 Wieviel Zeit benötigen Sie normalerweise für den Hinweg zur Arbeit oder Schule/Hochschule?
entfällt, da auf gleichem Grundstück
unter 15 Minuten
15 bis unter 30 Minuten
30 bis unter 60 Minuten
60 Minuten und mehr

Für Erwerbstätige (auch Selbständige, Landwirte, Mithelfende)

15 Geschäftszweig, (Wirtschaftszweig, Branche), zu dem der Betrieb (Firma, Dienststelle) gehört
Geschäftszweig des Betriebes

16 Sind Sie Arbeiter/Heimarbeiter
gewerblicher Lehrling
Angestellter
kaufm., techn., Verw.-Lehrling
Beamter/Richter
Selbständiger
Mithelfender Familienangehöriger

17 Wie lange arbeiten Sie normalerweise in der Woche?
unter 15 Std.
15-24 Std.
25-36 Std.
über 36 Std.

18 Üben Sie noch eine weitere Tätigkeit aus?
landwirtschaftliche Tätigkeit
sonstige Tätigkeit

BITTE NICHT AUSFÜLLEN

T 2 G.-D. 5 R. 6 St. 15 G.-D.

Für alle Personen

Wohnsitz am 1.9.1939 (Kriegsausbruch) Bundesgebiet einschl. Berlin (West) ☐ Sowjetzone bzw. Ostberlin ☐ Deutsche Ostgebiete ☐ Tschechoslowakei (einschl. Sudetenland) ☐ östliche Nachbarländer, Südosteuropa ☐ übrige Gebiete ☐

Für nach dem 1.9.1939 Geborene:
Wohnsitz des Vaters am 1.9.1939 oder, wenn Wohnsitz d. Vaters unbek., Wohnsitz der Mutter

Sind Sie nach Kriegsende aus der Sowjetzone bzw. aus Ostberlin zugezogen? ja ☐ nein ☐

Besitzen Sie einen Bundesvertriebenen- bzw. Bundesflüchtlings- ausweis? Ausweis A ☐ Ausweis B ☐ Ausweis C ☐

Wenn ja, welchen?

Für Nichterwerbstätige über 14 Jahre (auch Hausfrauen, Rentner, Pensionäre)

Waren Sie früher erwerbs- oder berufstätig? ja ☐ nein ☐

In welchem Jahr wurde die Erwerbstätigkeit aufgegeben? 19

Für Erwerbstätige (auch Selbständige, Landwirte, Mithelfende)

Ausgeübte Tätigkeit
a) Beruf (bei Beamten Amtsbezeichnung)

b) Stichwortartige Beschreibung (Arbeitsvorgang, Werkstoff, Ware u. dgl.)

Bedienen Sie bei Ihrer Tätigkeit überwiegend eine voll-/halbautomatische Maschine ☐ sonstige Maschine ☐ keine Maschine ☐

Wie hoch ist Ihr monatliches Nettoeinkommen aus Ihrer Erwerbstätigkeit? unter 300 DM ☐ 300 bis unter 500 DM ☐ 500 bis unter 800 DM ☐ 800 bis unter 1200 DM ☐ 1200 bis unter 1800 DM ☐ 1800 bis unter 2500 DM ☐ 2500 DM u. mehr ☐

(Für Selbständige in der Landwirtschaft und für Mithelfende Familienangehörige nicht zu beantworten.)

Für Personen in leitender oder aufsichtführender Stellung (ohne Selbständige)

Sind Sie Industriemeister, Werkmeister, Meister, Polier oder dgl. ☐ Kolonnenführer, Vorarbeiter oder dgl. ☐ Leiter, Geschäftsführer, Vorstandsmitglied oder dgl. eines Unternehmens od. Betriebes ☐ sonstwie leitend oder aufsichtführend tätig ☐

Für Selbständige

28 Wie viele Personen sind in Ihrem Betrieb beschäftigt? (einschl. tätige Inhaber, Leiter, Mithelfende Familienangeh.) Anzahl

29 Sind darunter Lohn- und Gehaltsempfänger? ja ☐ nein ☐

Für Besitzer, auch Pächter, von landwirtschaftlich genutzten Flächen. Bei Gartenland ab 0,5 ha (5000 qm) Vor Ausfüllung bitte Erläuterung genau lesen!

30 Wie groß ist die gesamte Fläche? (einschließlich gepachteter, aber ohne verpachtete Fläche) ha ar qm

Für Personen über 14 Jahre

31 Haben Sie eine praktische Berufsausbildung beendet? (Lehre, Anlernzeit oder dgl.) Wenn ja: Dauer in Jahren ja ☐ nein ☐

32 Auf welchen Beruf bezog sich diese Ausbildung?

Für Personen, die eine berufsbildende Schule bzw. Hochschul- ausbildung (auch Lehrerausbildung) abgeschlossen haben

33 Welche der folgenden Schulen haben Sie besucht und abgeschlossen? (nicht Berufsschule!) Technikerschule ☐ sonst. Berufsfach-/Fachschule ☐ Ingenieurschule ☐ Hochschule (auch Lehrerausbildung) ☐

34 Wie viele Jahre dauerte die gesamte Ausbildung an den in Frage (33) genannten Schulen bis zum letzten Abschluß? (siehe Erläuterungen) Anzahl der Jahre

35 Hauptfachrichtung des letzten Abschlusses? (siehe Erläuterungen)

36 Jahr des letzten Abschlusses? 19

Für Verheiratete, Verwitwete und Geschiedene

37 Wann wurde Ihre jetzige bzw. letzte Ehe geschlossen? 19

38 Waren Sie vorher schon einmal verheiratet? ja ☐ nein ☐

Für Frauen

39 Geburtsjahre aller lebendgeborenen ehelichen Kinder keine Kinder ☐

1. Kind 19	7. Kind 19
2. Kind 19	8. Kind 19
3. Kind 19	9. Kind 19
4. Kind 19	10. Kind 19
5. Kind 19	11. Kind 19
6. Kind 19	12. Kind 19

BITTE NICHT AUSFÜLLEN

A. d. ET 24 Ber. 24 A 28 Bach. 30 h a q 31 Abl. 32 E Ber. 34 D

Übersicht über die durch automatische Signierkontrollen
erkannten Fehler
 (Volkszählung 1961)

Merkmal	Fehler	
	Anzahl	in %
Haushaltsgröße	95	1,0
Geschlecht	731	7,5
Stellung zum Haushaltsvorstand ¹⁾	328	3,4
Geburtsjahr	85	0,9
Alterszuordnung (bezogen auf den Zählungstichtag)	1 194	12,3
Familienstand	573	5,9
Religionszugehörigkeit	2 933	30,2
Vorheriger Wohnsitz in der Sowjetischen Besatzungszone	275	2,8
Vertriebenen-, Flüchtlingsausweis	427	4,4
Erwerbs- oder Berufstätigkeit	278	2,9
Arbeitslosigkeit	429	4,4
Überwiegender Lebensunterhalt	841	8,7
Wochenarbeitszeit	138	1,4
Weitere Tätigkeit	88	0,9
Weg zur Arbeitsstätte	50	0,5
Schulart	148	1,5
Sonstige Merkmale	1 104	11,3
Insgesamt	9 717	100,0

1) Im Fragenkatalog 1970 wurde das Merkmal als "Stellung innerhalb des Haushalts" definiert.

Quelle: Storch, H., Zusammenfassender Bericht über die Erfahrungen bei der Volks- und Berufszählung 1961, Hrsg. Statistisches Landesamt Nordrhein-Westfalen, Düsseldorf 1965, S.69.

Übersicht über die durch automatische Kombinations-
kontrollen erkannten Fehler
 (Volkszählung 1961)

Merkmalskombination	Fehler	
	Anzahl	in %
Haushaltsgröße - entsprechende Ordnungsmerkmale	518	2,2
Personen unter 14 Jahre mit entsprechenden Merkmalskombinationen (z.B. ledig, nicht erwerbstätig)	3 276	13,7
Stellung zum Haushaltsvorstand/verheiratet/Alter	789	3,3
Ausbildung/Alter	150	0,6
Abgeschlossene Hochschulausbildung/Alter	179	0,7
Familienstand/Stellung zum Haushaltsvorstand	1 805	7,5
Männliche Personen unter 18 Jahre/erwerbs- oder nicht erwerbstätig	1 068	4,5
Schüler bzw. Student/Zeitaufwand und Verkehrsmittel (Weg zur Arbeitsstätte)	2 681	11,2
Eigener Lebensunterhalt/wirtschaftlich abhängig	2 711	11,3
Wirtschaftlich abhängig/fehlende Ernährerausgaben	3 138	13,1
Wirtschaftlich abhängig von einer Erwerbsperson/fehlende Ernährerausgaben	1 495	6,2
Überwiegender Lebensunterhalt/Ernährerausgaben	61	0,3
Zeitaufwand/Verkehrsmittel	189	0,8
Pendlerzielgemeinden/Zeitaufwand/Verkehrsmittel	757	3,2
Nichterwerbspersonen/Angaben über Erwerbstätigkeit (Wirtschaftsgruppe usw.)	2 235	9,3
100jährig/Stellung zum Haushaltsvorstand	6	0,0
Sonstige Merkmalskombinationen	2 910	12,1
Insgesamt	23 968	100,0

Quelle: Storch, H., Zusammenfassender Bericht über die Erfahrungen bei der Volks- und Berufszählung 1961, a.a.O., S.70.

Beispiele für personenbezogene deterministische Kontrollen

Wenn ...		Dann muß ...		Korrektur
Merkmal	Ausprägung	Merkmal	Ausprägung	
Schulbesuch und Alter	Typ A	Erwerbstätigkeit	nicht erwerbstätig	deterministisch
	Typ B	Schulbesuch	Volksschule	deterministisch
	Gymnasium und 5 bis 9 Jahre	Schulabschluß	leer	Angabe löschen
	Typ B Typ C	überwiegender Lebensunterhalt	Unterhalt durch Eltern eigenes Vermögen Sozialhilfe und sonst. Unterstützungen	Deck "Familienstand und Unterhalt nach Typen"
Ausgeübter Beruf	Typ D	Schulabschluß	Hochschule Fachschule leer Ingenieurschule leer Hochschule Technikerschule leer Fachschule leer Ingenieurschule leer mindestens 6 Jahre Medizin (Code 51) Geburtsjahr plus mindestens 26 Jahre	bei Ausbildungsdauer und Jahr des Schulabschlusses Mindestalter einsetzen; sonst Deck
	Arzt (Code 841)	Ausbildung (10 % - Teil)		
		Ausbildungsdauer Hauptfachrichtung Jahr des letzten Abschlusses		
Frühere Erwerbstätigkeit	Typ E ja	Jahr der Aufgabe der Erwerbstätigkeit	1920 bis 1970	Sonderdeck

Übersicht der wichtigsten Decks

Lfd. Nr.	Bezeichnung	Gespeicherte Merkmale je Deckzelle	Gliederung	Cold Deck
1	Alterseinsetzung a) nach Typen	Alter in Jahren	a) 5 Typen	Geburtsmonat 01 je Typ und Anstaltsart a) Alter in Jahren: Typ A-04, Typ B-10, Typ C-20, Typ D-35, Typ E-50 b) nach Kontrolle der Altersgrenzen für 40 Anstaltsarten je ein Alter, z.B. Säuglingsheim: 2 Jahre, Studentenwohnheim: 23 Jahre
2	b) nach Anstaltsarten		b) Anstaltsarten	
3	Einsetzung Familienstand, Unterhalt a) nach Typen	Familienstand, Unterhalt	a) 5 Typen	a) Typ A: ledig, Unterhalt durch Eltern Typ B: ledig, Unterhalt durch Eltern Typ C: ledig, Unterhalt durch Eltern Typ D: verheiratet, Erwerbstätigkeit Typ E: verheiratet, eigene Rente/Pension b) für 28 Anstaltsarten, z.B. Kinderdorf: ledig, Unterhalt durch Eltern, Altenheim: verwitwet, eigene Rente/Pension
4	b) nach Anstaltsarten		b) Anstaltsarten	
5	Personendeck	Lebensunterhalt Schulbesuch Schulabschluß Erwerbstätigkeit Geschäftszweig Stellung im Beruf Arbeitszeit Wohnsitz 1939 Zuzug (Frage 20) Frühere Erwerbstät. Maschinenbedienung Frühere Ehe	je männlich und weiblich nach Alter, Typ A: 0 bis 7 Typ B: 5 bis 14 Typen C,D,E: 15 bis 29, 30 bis 64, 65 u. älter	z.B. Typ D, 30 bis 64 Jahre, männlich: Lebensunterhalt: Erwerbstätigkeit Schulbesuch: - Schulabschluß: Volksschule Erwerbstätigkeit: erwerbstätig Geschäftszweig: Landwirtschaft Stellung im Beruf: Selbständiger Arbeitszeit: über 36 Stunden Wohnsitz 1939: Bundesgebiet Zuzug (Frage 20): nein Frühere Erwerbstätigkeit: - Maschinenbedienung: sonstige Maschine Frühere Ehe: nein

noch: Übersicht der wichtigsten Decks

Lfd. Nr.	Bezeichnung	Gespeicherte Merkmale je Deckzelle	Gliederung	Cold Deck		
				Ausbildungs- dauer:	Hauptfach- richtung:	Jahr des letzten Abschlusses:
6	Ergänzung zum Personendeck	Verkehrsmittel, Zeitaufwand	Typ B: 5 bis 14 Typ C,D: unter 18, 18 bis 64, 65 u. älter	Es wurde überwiegend "kein Verkehrsmittel" und als Zeitaufwand "unter 15 Minuten" eingesetzt. Die Korrektur war Voraussetzung für die Kontrolle der Pendelwanderung.		
7	Einsetzung Religion bei Insassen reli- giöser Anstalten	90%-Teil: einfacher, 10%-Teil: zusätzlich erweiterter Relig. code	keine (nur eine Anstaltsart)	römisch-katholisch		
8	Einsetzung Netto- Erwerbseinkommen nach Stellung im Beruf u. Geschlecht	Netto- Erwerbseinkommen	je männlich und weiblich Arbeiter Lehrlinge Angestellte Beamte Selbstständige (= Landwirt)	männlich: 800 - 1200 DM unter 300 DM 1200 - 1800 DM 1200 - 1800 DM 1800 - 2500 DM	weiblich: 500 - 800 DM unter 300 DM 800 - 1200 DM 1200 - 1800 DM 1800 - 2500 DM	
9	Einsetzung Berufs- bildungsmerkmale	Ausbildung (10%-Teil) Ausbildungsdauer Hauptfachrichtung Geburtsjahr	Techniker-, Fachschule Ingenieur-, Hochschule	2 Jahre 2 Jahre 3 Jahre 5 Jahre	Handels- schule Masch.bau höheres Lehramt	mind. 18 Jahre mind. 15 Jahre mind. 20 Jahre mind. 19 Jahre über Geburtsjahr

noch: Übersicht der wichtigsten Decks

Lfd.-Nr.	Bezeichnung	Gespeicherte Merkmale je Deckzelle	Gliederung	Cold Deck	
10	Einsetzung Beruf, Stellung im Beruf	Berufsbenennung, Stellung im Beruf	99 Wirtschaftszweige	z.B. Wirtschaftszweig: Kohlenbergbau Gießerei Bauhauptgewerbe Eh. mit Papierwaren und Druckerzeugnissen Straßenverkehr Wissenschaft, Bildung, Erziehung und Sport Organisationen des Wirtschaftslebens	Beruf: Bergmann Former Maurer Buchhändler Kraftfahrzeugführer Lehrer Geschäftsführer Angestellter
11	Einsetzung Wirtschaftszweigs, Stellung im Beruf	Berufsordnung, Wirtschaftszweig, Stellung im Beruf	328 Berufsordnungen	Die o.g. Kombinationen (Deck 10) finden sich in diesem Deck in anderer Reihenfolge wieder. Hier 2 zusätzliche Beispiele: Beruf: Wirtschaftszweig: Datenverarbeitungsfachmann Gebietskörperschaft Elektroingenieur Elektrotechnik	Stellung im Beruf: Angestellter Angestellter
12	Einsetzung Beruf, Wirtschaftszweigs, Stellung im Beruf nach Mindestalter	Berufsbenennung, Wirtschaftszweig, Stellung im Beruf	in Jahren: a) unter 21, b) 21 bis 24	Beruf: a) Verkäufer b) Schwerhörigenlehrer	Wirtschaftszweig: Eh. mit versch. Waren Wissenschaft, Bildung, Erziehung und Sport Angestellter

noch: Übersicht der wichtigsten Decks

Lfd. Nr.	Bezeichnung	Gespeicherte Merkmale je Deckzelle	Gliederung	Cold Deck
13	Haushaltsdeck	Stellung innerhalb des Haushalts bei der Mehrheit der Personen unter 18 Jahren Altersgruppe, Alter, Geschlecht, Familienstand, Stellung innerhalb des Haushalts, Lebensunterhalt, Typ	Vorkommen von Personen im Haushalt mit ... Jahren (Element eines Haushaltstyps) a) unter 18 (K) b) 18 bis 34 (J) c) 35 bis 50 (M) d) 51 und mehr (A) e) 18 bis 34 u/o 35 bis 50 (Ju/oM)	z.B. Haushaltstyp M + J oder M + J + K, Haushaltsvorstand: M. Kennzeichen für Stellung innerhalb des Haushalts bei der Mehrheit der Personen unter 18 Jahren: Sohn, Tochter des Haushaltsvorstands. Merkmal: Ausprägung bei 1. Person: 2. Person: 3. Person: M M J Altersgruppe M M J Alter 43 Jahre 41 Jahre 19 Jahre Geschlecht männlich weiblich männlich Familienstand verheiratet verheiratet ledig Stellung Ehegatte des Sohn des innerhalb des Haushalts- Haushalts- innerhalb des Haushalts- Haushalts- Haushalts- Haushalts vorstand vorstands vorstands vorstands vorstands Lebensunterhalt durch Erwerbs- Erwerbs- Erwerbs- Erwerbs- Erwerbs- tätigkeit D tätigkei E tätigkei D tätigkei D

Übersicht über die Lesefehler einer Arbeitsmasse
(70 382 Belege)

Lesefehler	Volkszählungsbogen	
	90 %	10 %
Rückweisungen	48	-
Kennziffer	15	-
Orientierungsmarke	16	-
Belegart	17	-
Geburtstag	69	10
Geburtsmonat	152	8
Substitutionen	70	4
Geburtsjahr	147	29
Sonstige Fehler im Geburtsdatum	109	9
Geschäftszweig	60	16
Pendelwanderung	32	4
Ausgeübte Tätigkeit	o.A.	21
Eheschließungsjahr	o.A.	13
Geburtsjahr der Kinder	o.A.	15
Sonstige Fehler des 10 %-Volkszählungsbogens	o.A.	22
Strichmarkierungsfehler	62	10
Verschmutzungen in Lesefeldern	14	3
Insgesamt	693	160

Lebenslauf

Am 25. August 1940 wurde ich in Großrückerswalde/Erzgebirge als Sohn des damaligen Segelfluglehrers Helmut Rönsch und seiner Ehefrau Marianne geb. Müller geboren.

Von 1946 bis 1950 besuchte ich dort die Volksschule. In Bad Münster/Deister setzte ich nach der Flucht meine Ausbildung an der Städtischen Mittelschule fort und wechselte 1952 zum Schiller-Gymnasium in Hameln. Dem Abitur folgte eine 2 1/2jährige Lehre zum Industriekaufmann bei den Continental Gummiwerken AG in Hannover. Das Studium der Volkswirtschaftslehre wurde 1962 an der TH Hannover aufgenommen und 1966 an der Universität Köln mit der Diplomprüfung abgeschlossen. Nach kurzer Tätigkeit in der Marktforschungsabteilung der Kodak AG in Stuttgart trat ich 1967 in das Statistische Landesamt Nordrhein-Westfalen in Düsseldorf ein, in dem ich von 1968 an als Dezernent zuständig war für die Volkszählung und seit 1974 mitverantwortlich für die Steuerstatistik.

Ab Oktober 1972 wurde mir ein Lehrauftrag an der Fachhochschule Düsseldorf erteilt. Im Februar 1977 wurde ich an die Fachhochschule für öffentliche Verwaltung Nordrhein-Westfalen in Düsseldorf abgeordnet.

Düsseldorf, den 20. Mai 1977

